Applied Intelligence (2024) 54:4370–4391
https://doi.org/10.1007/s10489-024-05399-w

Mediating effects of NLP-based parameters on the readability of crowdsourced wikipedia articles

Simran Setia[1] · Anamika Chhabra[2] · Amit Arjun Verma[3] · Akrati Saxena[4]

Accepted: 12 March 2024 / Published online: 26 March 2024

Abstract
In this era of information and communication technology, a large population relies on the Internet to gather information. One of the most popular information sources on the Internet is Wikipedia. Wikipedia is a free encyclopedia that provides a wide range of information to its users. However, there have been concerns about the readability of information on Wikipedia time and again. The readability of the text is defined as the ease of understanding the underlying text. Past studies have analyzed the readability of Wikipedia articles with the help of conventional readability metrics, such as the Flesch-Kincaid readability score and the Automatic Readability Index (ARI). Such metrics only consider the surface-level parameters, such as the number of words, sentences, and paragraphs in the text, to quantify the readability. However, the readability of the text must also take into account the quality of the text. In this study, we consider many new NLP-based parameters capturing the quality of the text, such as lexical diversity, semantic diversity, lexical complexity, and semantic complexity and analyze their impact on the readability of Wikipedia articles using artificial neural networks. Besides NLP parameters, the crowdsourced parameters also affect the readability, and therefore, we also analyze the impact of crowdsourced parameters and observe that the crowdsourced parameters not only influence the readability scores but also affect the NLP parameters of the text. Additionally, we investigate the mediating effect of NLP parameters that connect the crowdsourced parameters to the readability of the text. The results show that the impact of crowdsourced parameters on readability is partially due to the profound effect of NLP-based parameters.

Keywords Crowdsourcing · Readability · Lexical Diversity · Mediation analysis

1 Introduction

In Human-Computer Interaction (HCI), the readability of the text is one of the primary requirements for HCI systems designed to provide textual information [1]. The readability of the text is defined as the degree of comprehension imparted by the underlying text. One such HCI system serving textual information is Wikipedia, which is regarded as one of the most exhaustive sources of textual information. According to Alexa rankings, Wikipedia has outperformed all other information sources on the Internet with respect to users' traffic [2]. Wikipedia is a crowdsourced information source curated and maintained by the crowd. Any Wikipedia article can be edited by any individual called the editor. The edits done by the editors include addition, deletion, rephrasing, and restructuring the given information present in the article. Hence, the crowdsourced nature of Wikipedia articles allows the union of the knowledge of different people around the globe. Given that Wikipedia is a source of exhaustive information, the majority of Wikipedia users are readers rather than producers of information [3]. The high traffic on Wikipedia comprises mostly readers reading the information rather than editors who contribute to the Wikipedia articles. For a reader, one of the primary requirements is the comprehension of the underlying text, which is captured by readability. In view of the importance of readability for an HCI system serving textual information and the popularity of Wikipedia as an information source, it becomes imperative to study the readability of Wikipedia.

1.1 Research gap

According to one of the previous studies on the readability of Wikipedia, it is established that the crowdsourced

✉ Akrati Saxena
a.saxena@liacs.leidenuniv.nl

Extended author information available on the last page of the article

nature of Wikipedia articles does affect their readability [4]. The crowdsourced parameters used to capture the nature of crowdsourced articles include the edit coefficient, stylistic coefficient, and knowledge gap parameter. In [4], the authors develop a classification model to distinguish between readable and non-readable articles, and conclude that the knowledge gap parameter is the most important parameter that affects readability. The knowledge gap parameter captures the missing pieces of information in Wikipedia articles, which hinder the comprehension of the underlying text. It should also be noted that many of the past NLP-based studies have shown a number of text-based parameters, such as lexical diversity, semantic diversity, lexical complexity, semantic complexity, are correlated with the readability of the underlying text [5–10]. However, there is no such study which analyses the effect of all possible NLP based parameters on the readability of text. All of the aforementioned studies study the effect of one or two parameters on the readability at a time. In addition to this, there is no such study on crowdsourced platforms such as Wikipedia that analyses the effect of NLP parameters and crowdsourced parameters on the readability of articles. The present study explains the effect of both NLP parameters and crowdsourced parameters on the readability of the articles. The present study also takes into consideration the change in NLP parameters due to crowdsourced parameters, which in turn affects the readability of the text.

It is clear from the aforementioned studies that the text-based parameters affect the readability of the text. However, in the case of Wikipedia articles, the crowdsourced parameters also affect the readability of Wikipedia articles. In the present study, we first use existing tools and techniques to calculate the NLP-based features. We then build an Artificial Neural Network (ANN) to quantify the effect of NLP-based features on the readability of articles. According to the results obtained, the NLP-based features affect the readability of the articles. However, it should be noted that it is the crowd that influences these features in the case of Wikipedia articles. Since any piece of text can be added, deleted or modified by the editors of the article on Wikipedia, multiple editors author a single Wikipedia article. Hence, we can say that the crowd dynamics on Wikipedia is responsible for the changes in NLP-based features of the article, which in turn lead to changes in the readability scores. To quantify the effect of crowdsourced parameters on the readability of articles, we do not only analyze the direct effect of crowdsourced parameters on readability scores, but we also analyze the effect of crowdsourced parameters on NLP-based features that lead to changes in readability scores. Hence, we perform mediation analysis with NLP-based features as the mediating variable to study the effect of crowdsourced parameters on readability scores. The mediation analysis helps us to study the effect of crowdsourced parameters on readability scores, particularly due to the changes in NLP-based features. The aim of mediation analysis is to quantify the extent to which the effect of crowdsourced parameters on readability can be modified through NLP-based features. The results show that the effect of crowdsourced parameters on the readability of articles is partially due to the changes in NLP-based features introduced by the crowd.

1.2 Present study

In this study, we introduce five NLP-based parameters related to the readability of the text. The parameters considered are semantic complexity, lexical complexity, sentiment analysis, lexical diversity, and semantic diversity exhibited by the text. These parameters are selected as they have been found to be correlated with the readability of the text [11, 12]. We train a feed-forward artificial neural network using these parameters to study the relation between these parameters and the readability of the articles.

However, the crowd dynamics may be responsible for achieving the desired text-based features like lexical diversity. There is a high probability of the text being more lexically diverse if it is written by multiple editors. The footprints left by the crowd/editors while editing the readable article must be discerned to engineer the production of more readable articles in Wikipedia. We investigate the various parameters specific to the crowdsourced nature of Wikipedia articles that may affect the text-based features, which in turn affects the readability. The crowdsourced parameters investigated are the ratio of experienced editors editing the particular article and the standard deviation of edits contributed by the editors to the articles. For this, we design a theoretical model that judges the mediating effects of NLP-based features (particularly lexical diversity) between the above-mentioned crowdsourced parameters and the readability of Wikipedia articles. The mediation analysis judges whether the effect of crowdsourced parameters on readability can be mediated by the change in NLP-based features or not.

The present study is divided into two phases. In the first phase, we build an artificial neural network to find out the most dominant NLP-based feature affecting the readability of articles. According to the results obtained, the parameters related to lexical diversity are found to be the most dominant parameters that affect the readability of Wikipedia articles. In the second phase, we perform mediation analysis, which judges whether the effect of crowdsourced parameters on readability can be modified by NLP-based features or not. The results of the mediation analysis suggest that a partial mediation effect of lexical diversity exists between the two crowdsourced parameters and the readability of the Wikipedia articles. Hence, the first phase verifies the effect of NLP-based parameters on the readability of Wikipedia articles. The second phase judges the effect of crowdsourced parameters on the readability of articles. Further, we perform

a mediation analysis that verifies the effect of crowdsourced parameters on readability scores due to lexical diversity as the mediator variable. The mediation analysis sheds light on the effect of crowdsourced parameters on lexical diversity, which in turn affects the readability scores.

Overall, the findings suggest the importance of including text-based features in studying the effect of crowdsourced parameters on the readability of Wikipedia. The main contributions of our work are as follows.

1. Understanding the impact of NLP-based features on the readability of crowdsourced articles.
2. Understanding the impact of crowdsourced parameters on the readability of articles.
3. Analyzing the effect of crowdsourced parameters on lexical diversity of articles.
4. Finding out the mediating effect of NLP-based parameters on readability due to the crowdsourced parameters.

1.3 Organization of the article

The remainder of the article is organized as follows: Section 2 discusses the related studies on NLP-based features affecting the readability of articles. A summary of recent studies on the readability of Wikipedia articles is also discussed in this section. Then, dataset details and tools used are discussed in Section 4 and Appendix 6, respectively. Next, we divide the methodology and results sections into two phases. Section 5.1 discusses the methodology and results obtained for the neural network built based on NLP-based features (Phase 1). Section 5.2 discusses the methodology and results of mediation analysis of NLP-based features on readability due to crowdsourced parameters (Phase 2).

2 Related work

In our study, we consider NLP-based parameters and crowdsourced parameters affecting the readability of Wikipedia articles. In this section, we first discuss the past studies based on NLP parameters that affect the readability of a text article. Next, we discuss the literature based on crowdsourced parameters that specifically affect the readability of Wikipedia articles.

2.1 NLP Based parameters

This section discusses the literature based on NLP-based parameters and their effect on the readability of the underlying text. It should be noted that this section enumerates all the past studies that describe the NLP-based parameters affecting the readability of text. All the NLP-based parameters can be categorized under five broad categories, which are listed below.

2.1.1 Cohesion

Many of the past researchers have investigated several parameters that can act as predictors for text readability. One of such parameters is the cohesion of the text. Cohesion refers to grammatical linking in the sentences, which imparts meaning to them. Past studies have clearly shown that cohesive texts are very easy to comprehend for a reader [13–16]. In 2013, Todirascu et al. [6] demonstrated that the cohesion of the text is a predictive variable for the readability of the text. The authors calculated the correlation scores of the readability of the text (annotated manually) with 41 variables measuring the cohesion of the text. Some of the variables were calculated using similarity between the adjacent sentences using cosine similarity in LSA[1] space, POS[2] tagging, and many other techniques. LSA-based parameters were found to be highly correlated with the readability of the text.

In addition to this, one more study conducted by Rezeaee et al. [7] showed the correlation between readability and cohesion. In this particular study, readability was calculated using standard readability metrics, and cohesion was calculated using grammatical markers, lexical markers, and conjunctions such as and, or, then, and so on. In another study focused on calculating concept-based readability, document cohesion was used as one of the parameters to calculate the desired readability metric [17]. Document cohesion can be calculated using Leacock-Chodorow semantic similarity measure [18]. Document cohesion captures the cohesion between concepts explained in the documents. The results show that more document cohesion led to better readability values.

2.1.2 Sentiment analysis

The sentiments conveyed by the text are associated with the feelings, emotions, and opinions subjected by the text [8, 19]. The readability of text is also found to be a function of the sentiments conveyed by the text [20]. This is because a reader's opinion may differ from the one conveyed by the text. Also, some of the past studies claim that the opinion conveyed in the text influences cognition and, hence, the comprehension of the particular text [21].

There are generally two ways of measuring the underlying sentiment [22]. One is a lexical methodology, where a list of sentiment words (such as good, bad, happy) along with the intensity of the sentiment conveyed by the word

[1] Latent Semantic Analysis

[2] Parts-of-Speech

are already given. We can quantify the sentiment of a sentence by extracting the sentiment words and then using the given list. Another way of quantifying the sentiment is Bag-Of-Words, which uses techniques such as POS tagging and co-occurrence with other sentiment words.

There are a number of studies that quantify or classify the sentiment conveyed by the underlying sentence. One of the past studies employs a hybrid approach to classify the sentiment (positive/negative) conveyed by the sentence [23]. The authors extract sentiment words using POS Tagging and LDA [3] Topic Modelling. After the extraction of words, they employ AFINN (lexicon-based dictionary) to capture the context of the underlying sentiment word. This approach helps to classify ambiguous words with high accuracy. In addition to this, a number of deep learning models, such as the BERT model, are also used to quantify the sentiment conveyed by the text [24]. BERT is used to generate text embeddings, which can then be used as features in the classification model to classify the sentiment conveyed by the underlying text. Another study focuses on capturing the variance in sentiments conveyed by the underlying sentiments [25]. The authors construct a novel feature set that helps quantify the variance in the intensity of the sentiments.

2.1.3 Lexical complexity

Apart from the parameters discussed above, lexical complexity also affects the readability of the text. Cobb et al. [11] observed that text comprehension is dependent on the words presented in the underlying text. If the reader is aware of the words used in the text, then the comprehension of the text by the reader is successful. In addition to this, Crossley et al. [9] also claimed lexical complexity to be a determinant in calculating the readability of the text. However, the lexical complexity was only calculated in terms of the frequency of the words present in the text. A class of studies also focused on simplification of the underlying text to make it more readable [9, 26]. Simplification involves rephrasing the complex lexical as well as syntactic structures, which aids the readers in the comprehension of the text.

Lexical sophistication is also considered one of the dimensions of lexical complexity. Various past studies have established that the measures used to quantify lexical sophistication, such as word familiarity [27], word imageability [28], and word concreteness [28], affect the readability of the text.

2.1.4 Syntactic and semantic complexity

Similar to lexical complexity, syntactic complexity is also found to be an indicator of the text readability [12]. A past study conducted an experiment with three different versions of a text with varied syntactic complexities. The readers were also divided into three groups based on their reading proficiency. The results suggested that syntactically complex texts were difficult to comprehend in case of low proficient readers. In addition to this, a past study has concluded that a semantically complex sentence is difficult to comprehend, and hence, the readability of the semantically complex text is always low [29].

The syntactic complexity of the underlying sentence is measured using metrics, such as mean length of sentences, mean length of clauses, and mean length of phrases [30]. The rationale behind using these simple metrics is that a longer sentence is syntactically more complex as compared to a shorter sentence. On the other hand, the semantic complexity can be measured with the help of text embeddings using BERT [31]. The semantics of the underlying text can also be discerned with the help of topological embeddings using a graph based on entities present in the text [32].

2.1.5 Lexical diversity

Lexical diversity is termed as one of the factors that affects the readability of the underlying text [5]. If the underlying text is lexically diverse, that means it uses more number of unique words. It can also be said that it tries to convey more information in less amount of text. In such cases, it becomes difficult to comprehend the underlying text, and the readability is low. In another study regarding the readability of Journal Mission Statements (JMSs), it was established that the readability and lexical diversity were positively correlated with each other [33]. In [34], the authors developed a new readability measure called CAREC (Crowdsourced Algorithm of Reading Comprehension) that uses crowdsourcing techniques to collect human judgments of text comprehension. This study used lexical diversity and some other features, including surface-level parameters (including the number of words, number of phrases, and number of sentences), lexical sophistication, and so on, to predict the readability scores.

There are a number of past studies that have coined indices to measure the lexical diversity of the text [35, 36]. Some of the important indices are evenness, i.e. the variance in words observed for different types; dispersion, i.e. the average number of words between the words of the same type; and importance, i.e. the frequency of the words in the text as a whole. These indices help measure the lexical diversity of the text.

2.2 Crowdsourced parameters

A few past studies have worked on the readability of crowdsourced articles [4, 37, 38]. Lucassen et al. [37] calculated the readability scores of Wikipedia articles using conventional

[3] Latent Dirichlet Allocation

readability metrics like the Flesch Kincaid Readability Score. The results of the study suggested that Wikipedia articles are hard to comprehend since they score low on readability. However, it should be noted that the conventional readability metrics are based on surface-level metrics, such as the number of words in a sentence, the number of sentences in a paragraph, the number of phrases in a sentence, etc. Such metrics do not take into account the semantics of the text and, hence, are not sufficient to quantify the readability of the text [39, 40]. Semantic text mining is an important dimension required to quantify the readability of the text [41, 42]. Keeping in mind the shortcomings of the study, some other studies defined some metrics that capture the semantics/quality of the Wikipedia articles. In one such study, the authors built a classification model to classify Wikipedia articles into readable and non-readable articles [4]. The authors considered NLP features and crowdsourced features to classify the given set of articles. The crowdsourced features considered were the edit coefficient, stylistic coefficient, and knowledge gap parameter [4]. In addition to the above studies, a number of studies have been conducted in the direction of analyzing the readability of medical information present on Wikipedia [43–46]. [47] focused on the analysis of patient medication information and showed that the readability scores of the information given on Wikipedia were usually higher for an average reader to understand. In [38], the authors compared the readability of Wikipedia articles with the Simple Wikipedia articles (Simple Wikipedia is one of the projects by Mediawiki foundation that maintains concise and readable versions of Wikipedia articles) and Britannica articles using parameters related to word complexity in conjunction with conventional readability metrics. The authors scrutinized the words present in the text based on topic-based familiarity, genre-based familiarity, and popularity-based familiarity. The results show that Wikipedia articles score low on readability as compared to Simple Wikipedia and Britannica. To summarize, there are a number of crowdsourced parameters defined that affect the readability of the text, including the knowledge gap parameter, stylistic coefficient, edit coefficient, and word familiarity. As per the above discussion, there are two types of parameters (NLP-based/crowdsourced) that lead to changes in the readability scores of an article. In the case of NLP parameters, the high-level metrics are cohesion, lexical diversity, sentiments, lexical complexity, syntactic complexity, and semantic complexity [48]. These metrics are, in turn, measured using a number of low-level metrics. Similarly, in the case of crowdsourced parameters, the high-level metrics are edit coefficient, stylistic coefficient, and knowledge gap parameters, and a number of low-level metrics help to measure these high-level metrics. The summary of all metrics described in this section is provided in the Table 1.

3 Tools and techniques

The conventional readability metrics only consider surface-level parameters such as the number of words, sentences, and phrases to calculate the text readability, which do not capture the quality of the written text [40, 51]. Therefore, we use a number of text analysis tools, including TAALED, SEANCE, TAALES, TAASSC, and TAACO, to compute quality parameters related to the readability of the text, such as lexical diversity, lexical complexity, and semantic complexity. These tools are briefly discussed below. We further summarize the mapping of high-level metrics to low-level metrics in Table 2.

3.1 TAALED

TAALED (Tool for Automatic Analysis of Lexical Diversity) [52] is used to measure the lexical diversity of a given text using three dimensions, i.e., Volume, Abundance, and Variety. Volume, Abundance, and Variety take into account the number of tokens, the number of different types of tokens, and the number of different types of tokens encountered in a given length of the text, respectively. TAALED computes a number of indices to quantify these three dimensions of lexical diversity, and further details are provided in Appendix A.

3.2 TAASSC

TAASSC (Tool for Automatic Analysis of Syntactic Sophistication and Complexity) is used to measure various indices related to syntactic complexity and syntactic sophistication. The syntactic complexity is measured using phrasal complexity and clausal complexity. It should be noted that clauses are different from phrases in English grammar. A clause must have a subject and a predicate, whereas a phrase does not have a subject and a predicate. On the other hand, syntactic sophistication is measured using a number of frequency-based indices like the frequency of verbs and the conditional probability of verbs occurring in a specific sentence structure (for example, a sentence comprising of subject, verb, and object). A detailed explanation of indices is given in Appendix A.

3.3 TAALES

TAALES (Tool for Automatic Analysis of Lexical Sophistication) is used to measure indices related to lexical sophistication [53]. Lexical sophistication of the underlying text takes into account the length and breadth of lexical words, i.e., the frequency and the quality of the words used in the text. The frequency is calculated using the frequency of dif-

Table 1 Past Literature on NLP-based parameters and Crowdsourced Parameters

Past Study	High Level Metric	**Low Level Metrics**	Type of Parameter (NLP/Crowdsourced)
Todirascu et al. [6]	Cohesion	Similarity between sentences using overlap in POS Tags, cosine similarity in LSA space, overlap between subject and object of the sentences	NLP
Rezeaee et al. [7]	Cohesion	Number of grammatical markers, conjunctions, and lexical markers	NLP
Yan et al. [17]	Cohesion	Leacock-Chodorow semantic similarity	NLP
Shapiro et al. [22]	Sentiment	Predefined lexical list of words and BOW methodology	NLP
Crossley et al. [9]	Lexical Complexity	Finding complex words using word familiarity [27], word imageability [28], and word concreteness [28]	NLP
Lu et al. [30]	Syntactic Complexity	Mean length of sentences, clauses, and phrases	NLP
Jarvis et al. [49]	Lexical Diversity	Dispersion, Importance and Evenness of the words	NLP
Ren et al. [50]	Edit Coefficient	Number of edits done on a Wikipedia article	Crowdsourced
Jatowt et al. [38]	Word Familiarity	Topic based familiarity and Popularity of word	Crowdsourced
Setia et al. [4]	Stylistic Coefficient	Coefficient of variance in number of editors editing different sections of Wikipedia article	Crowdsourced
Setia et al. [4]	Knowledge Gap Parameter	Semantic similarity between sentences using LDA.	Crowdsourced

ferent words used and the n-gram frequency. On the other hand, the quality of the words is measured using the indices like correctness, familiarity, and imageability. The description of all these indices is given in Appendix A.

3.4 TAACO

TAACO (Tool for Automatic Analysis of Cohesion) is used to measure the indices required for text cohesion. The text cohesion is measured through lexical overlap and semantic overlap [54]. The lexical overlap between sentences is measured by the number of intersecting POS tags between the sentences, and the lexical overlap between paragraphs is measured by the number of intersecting POS tags between various sentences or paragraphs. On the other hand, semantic overlap between sentences is measured using the number of semantically similar words present between sentences. Similarly, in the case of paragraphs, semantically similar words are found between sentences of given paragraphs. It should be noted that semantically similar words are found using WordNet. The information about all these indices is given in Appendix A.

Table 2 Mapping of high-level metrics to low-level metrics

Tools	High Level Metrics	Low Level Metrics
TAALED	Lexical Diversity	Volume, Abundance, and Variety
TAASSC	Syntactic Complexity	Phrasal and Clausal Complexity
TAASSC	Syntactic Sophistication	Frequency of Verbs and Conditional Probability of Verbs
TAALES	Lexical Sophistication	Word Frequency, n-gram Frequency, Correctness, Familiarity, Imageability
TAACO	Text Cohesion	Lexical and Semantic Overlap
SEANCE	Sentiment Score	Frequency of positive and negative words using preexisting databases

Table 3 Mean and Standard Deviation of four rating dimensions present in AFT

Measures	Trustworthiness	Neutrality	Completeness	Readability
Mean	3.80	3.76	3.62	3.91
Stdev	1.45	1.46	1.47	1.39

3.5 SEANCE

SEANCE (Sentiment Analysis and Social Cognition Engine) is used to extract indices related to human sentiments and feelings conveyed by the underlying text. The sentiments are quantified using preexisting databases that define the degree of positive and negative sentiment conveyed by a particular word. The preexisting databases are SenticNet [55, 56], VADER [57], Hi-Liu polarity [58], Emolex [59]. More details about the indices calculated are given in Appendix A.

4 Dataset

The dataset is collected using Article Feedback Tool (version 4) (AFT) and is publicly available on Wikipedia data dump [60]. Article Feedback Tool is a tool deployed by the Mediawiki Foundation on Wikipedia articles to collect feedback from the readers. The main aim of this initiative is to engage readers and assess the quality of Wikipedia articles. AFT was deployed on around 45000 English Wikipedia articles. AFT assisted the readers to rate the Wikipedia articles on four dimensions, namely trustworthiness, completeness, neutrality, and readability [61]. All the dimensions were rated on a scale of one to five. The dataset consists of 12,498 ratings given by various users. The statistics related to the four rating dimensions are given in Table 3. The distribution of four rating dimensions is also given in Fig.1. The frequency distributions for all the rating dimensions are observed to be skewed. In order to carry out the experiments, we sampled 3000 articles out of 45000 articles. We use stratified sampling to sample these articles. As shown in Fig. 1, we have articles

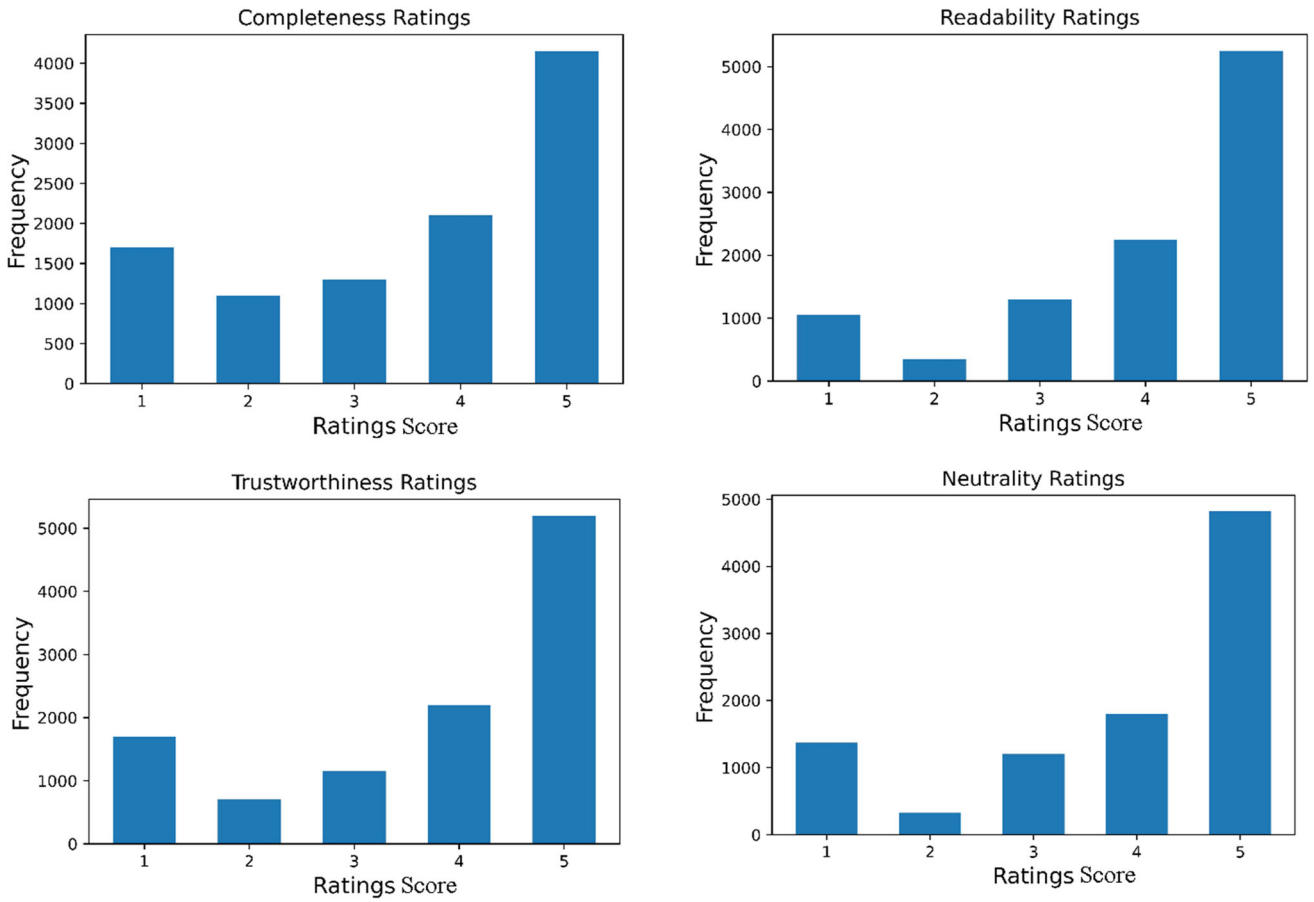


Fig. 1 Frequency Distribution of four rating dimensions of AFT. The above figure shows that all the rating dimensions follow a skewed distribution

rated on a scale of one to five. So, we divide the given set of articles into five strata according to the ratings and sample an equal number of articles (i.e., 600) from each strata.

NLP-based features are collected using the TAALED, TAACO, TAASC, TAALES, and SEANCE that quantify lexical diversity, cohesion, semantic complexity, lexical sophistication, and sentiments in text, respectively. The details of the features are as follows.

- TAALED is used to calculate 38 indices for measuring lexical diversity.
- TAALES is used to calculate 484 indices for measuring lexical complexity.
- TAASSC is the tool used to quantify 367 indices related to syntactic complexity. It is used to calculate 31 indices related to clausal complexity, 132 indices related to phrasal complexity, 190 indices related to syntactic sophistication, and 14 indices included in Syntactic Complexity Analyzer.
- TAACO is the tool used to calculate 178 indices related to text cohesion, out of which 15 indices are related TTRs, 54 indices are related to the lexical overlap of sentences, 54 indices are related to the lexical overlap of paragraphs, 8 indices are related semantic overlap of sentences, 8 indices are related to the semantic overlap of paragraphs, 25 indices are related connectiveness, and 4 indices are related to givenness.
- SEANCE is the tool used to calculate 254 indices measuring the sentiments conveyed by the text.

5 Experimental analysis

In this section, we will discuss the experiments carried out as a part of the research study. We build a neural network to quantify the readability scores. The neural network helps to find out the most dominant NLP feature that affects the readability of articles. The most dominant NLP feature is then used in mediation analysis to find out the mediating effect of NLP features in the relationship between crowdsourced parameters and the readability of articles.

5.1 Experiment 1: deep learning based approach to quantify readability

In the first phase of the research study, we use a feed-forward Artificial Neural Network (ANN) to quantify the readability scores based on input NLP features computed using tools, as discussed above.

5.1.1 Details of experiment

We use a neural network to quantify different readability scores of the articles. We use NLP-based parameters as the features and four dimensions (trustworthiness, completeness, neutrality, and readability) present in the dataset as the target variables. For this, we perform text preprocessing on 3000 Wikipedia articles present in the dataset. The text preprocessing of Wikipedia articles involved removing the unwanted words and tags used in the Wiki markup language. After performing the text preprocessing, the text corpus of 3000 articles is fed to various tools (TAALED, TAASSC, TAALES, TAACO, AND SEANCE), which calculate 1474 indices used as input to the neural network, as shown in Fig. 3. The regression models are built using 1474 independent variables and four respective dependent variables (Readability, Trustworthiness, Neutrality, and Completeness), as discussed in the dataset. We take all four survey parameters as the dependent variables to design the regression model aimed at quantifying the readability scores. It should be noted that the three survey parameters other than readability are also essential for an article to achieve the desired readability score. An incomplete article contains missing information, which hinders the comprehension of the underlying text [4]. An article written from a specific point of view may not be appreciated by a reader from a contrasting point of view. Some of the past studies suggest that the opinion conveyed by the text influences the cognition and comprehension of the underlying text [21]. Therefore, an article must be written from a neutral point of view. In addition, the reliability of the information present in the text affects the readability of the text intuitively. If a reader does not rely on the information conveyed by the text, then he/she will not certainly make any effort to comprehend the information given in the text. Hence, we use all four survey parameters as the dependent variables to quantify the readability scores. The proposed methodology of experiment 1 is shown in Fig.2.

We train a feed forward neural network with three backpropagation training algorithms. We use a train-test split of 80:20. The neural network comprises of one input layer, three hidden layers, and one output layer for each of the four target variables. We use three training algorithms to train the neural network, which are listed below.

1. Scaled Conjugate Gradient Algorithm (SCG) [62]
2. Levenberg-Marquardt Algorithm (LM) [63]
3. Gradient Descent Algorithm [64]

As per the results obtained, we observe better mean accuracy with Gradient Descent Algorithm (see 5). The input layer

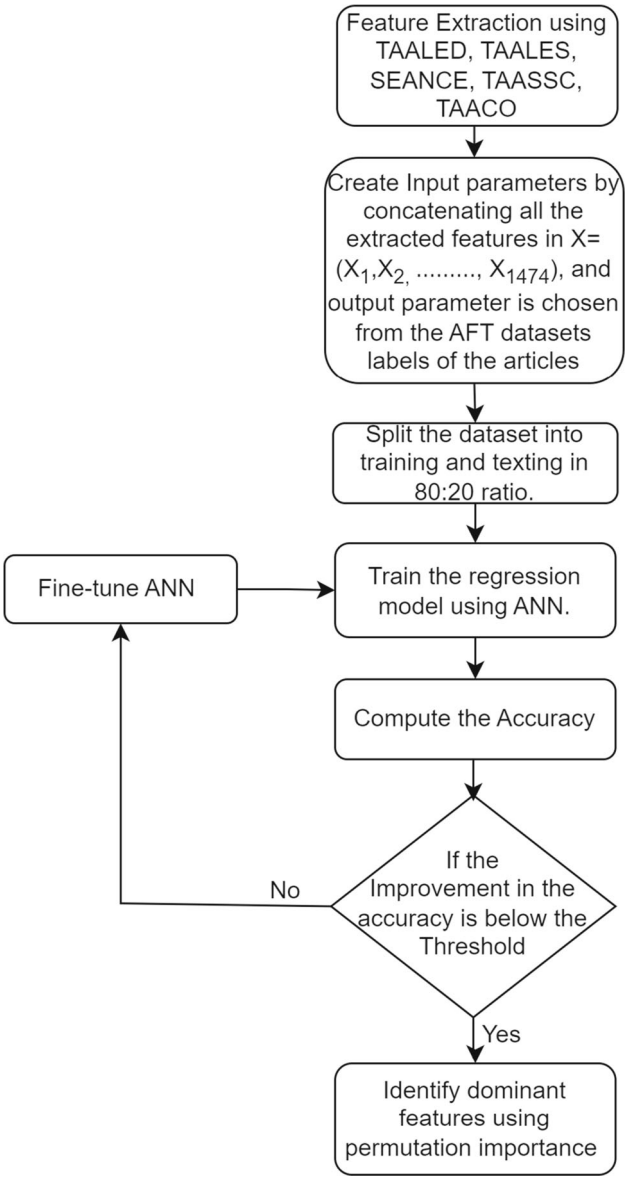


Fig. 2 Phase 1 Methodology: This flowchart describes the steps to be followed to execute experiment 1. First, we collect the features using the tools mentioned in the previous section. Then, we build four regression models using the features as independent variables and each of the AFT dimensions as the dependent variables. We fine-tune the hyperparameters of the ANN by observing the changes in the accuracy of the regression models

comprises 1474 neurons for processing all the inputs. Each of the hidden layers comprises 3000 neurons, and the output layer has four neurons, which output the readability, trustworthiness, neutrality, and completeness scores (as shown in Fig. 3). In addition to this, we use the RELU activation function in the output layers, and we use the hyperbolic tangent function for the hidden layers. The performance goal accuracy is set to 85% as we observe no significant improvement in accuracy after achieving the performance goal, and the improvement is below the threshold (less than 0.5%). After tuning the hyperparameters, the learning rate is 0.05. The number of epochs is set to a high value of 3000, and the batch size is 100. The number of epochs required for convergence for different training algorithms and dependent variables are listed in Table 4. The training of the neural network stops once it reaches the required performance goal. The training and testing accuracies of all four models with three different training approaches are listed in Table 5. The designed ANN model exhibited an accuracy of 84.7% with readability as the dependent variable, 87.2% with trustworthiness as the dependent variable, 83.2% with neutrality as the dependent variable, and 86.3% with completeness as the dependent variable. The mean accuracy observed is 85.4%. The accuracy is calculated using the difference between the actual and predicted values and then subtracting the difference from 1. The hyperparameters are tuned using the Grid Search method of hyperparameter optimization. A number of alternatives are tried for each of the hyperparameters, and those that exhibit higher accuracy are selected. It should be noted that the GridSearchCV method in Keras uses predefined splits for training and testing along with cross-validation [65]. The results shown in Table 5 are for specific train and test splits when the GridSearchCV method is overridden with a predefined split. The predefined train and test split used is 80:20.

5.1.2 Results of experiment 1

If we talk about the relative importance of the features involved, lexical diversity is the most dominant feature observed in the entire feature space. Since lexical diversity is associated with a number of indices, we observe that the index *mtld_ma_wrap_aw* is found to be the most dominant one. The mtld_ma_wrap_aw index is computed using TAALED and is based on the moving average of content words to reach a certain Type Token Ratio (TTR) value. The TTR is the ratio of types of tokens to the total number of tokens occurring in a text [52]. To identify the dominant feature, we use the permutation importance method to find the importance of features in ANN [66]. The permutation importance technique permutes the given input feature and checks the change in accuracy. If the accuracy drops, it means that the particular feature affects the output parameters, and if it does not drop, it means the feature under consideration does not have any impact on the output features. Based on the drop in accuracy, the permutation importance technique ranks the features. The relative drop in ANN accuracy observed for all the features is mentioned in the table given in Appendix B. Here, the drop in accuracy refers to the drop in mean accuracy. It should also be noted that the second highest drop in accuracy is observed for a feature which is also related to lexical diversity, i.e., mltd_ma_bi_aw. mltd_ma_bi_aw is based on the

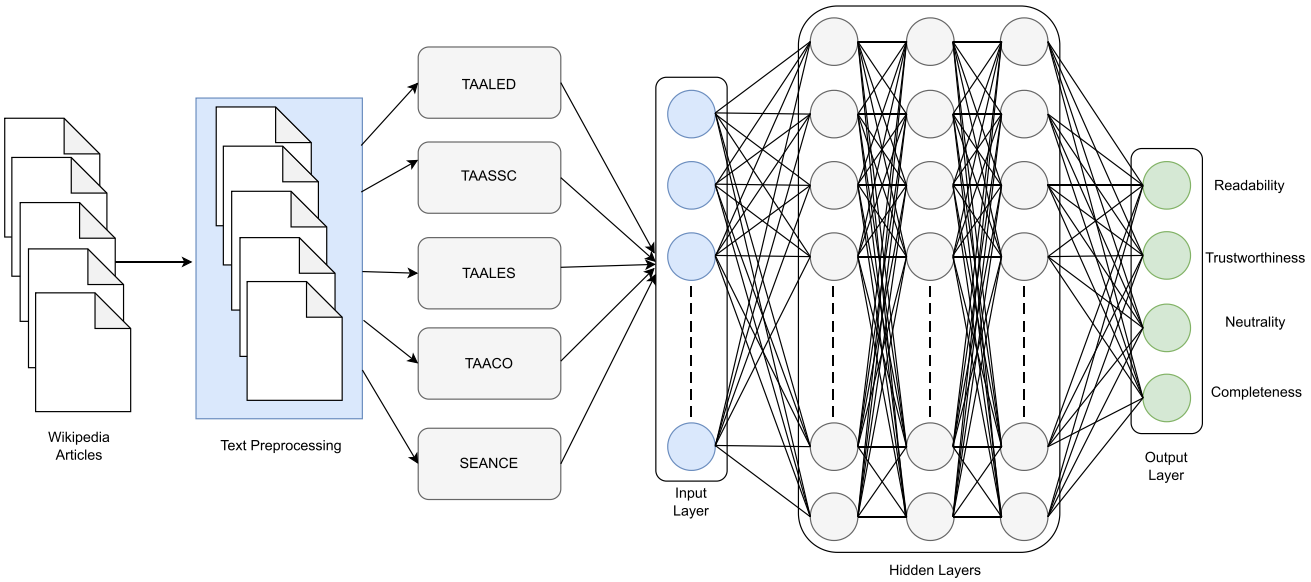


Fig. 3 Workflow for ANN-based approach to compute readability. The above figure depicts the phase 1 of the underlying analysis. First, we extract the features of each article using tools such as TAALED, TAASSC, TAALES, TAACO, and SEANCE. Once the features were extracted, we built ANN with these features as the independent variables and each of the rating dimensions as the dependent variables

moving average of content words both in forward and backwards directions to reach a certain Type Token Ratio (TTR) value.

Next, we plot the most dominant feature against all the target variables (Fig. 4). We observe a positive correlation between lexical diversity (mtld_ma_wrap_aw) and all the target variables. The positive correlation shows that a higher value of lexical diversity facilitates better readability scores. A higher value of mtld_ma_wrap means more words are required to reach the TTR value. This, in turn, implies that fewer diverse tokens are used in the text. If fewer diverse tokens are used, that means the reader encounters very few new words while reading the text. Fewer new words clearly indicate that the readers can easily comprehend the underlying text. This is also proved by one of the past studies about readability. The study states that if more diverse tokens are used, then it is difficult for a reader to comprehend the underlying text [5]. However, fewer diverse tokens facilitate comprehension of the text and, hence, better readability scores.

In addition to this, the other parameters (Trustworthiness, Neutrality, and Completeness) also exhibit a positive correlation with lexical diversity. The possible reasoning behind this is the positive correlation of all of these parameters with the number of edits. Trustworthiness is measured by a number of references in the article. As number of edits increases in the article, number of references also increases. An article also comes close to the completion stage with a number of edits. The neutrality of the article also increases if there are more edits done by different editors. A single editor editing the article may express only his point of view. In contrast to this, when a number of editors express their point of view, it leads to a neutral point of view. It should also be noted that the lexical diversity increases with more number of edits. Hence, as the number of edits increases in the article, lexical diversity increases and the parameters (Trustworthiness, Neutrality, and Completeness) also increase.

Further, it should be noted that the kind of tokens used in the text affects the readability of the underlying text. If more diverse tokens are used, then the readability is low. However, if the diverse tokens used are content words[4], such as I, we, are, is, you, and many more, then it should not affect the readability. In case the diverse tokens refer to function words[5], such as nouns, adjectives, and adverbs, then it must affect the readability. In order to verify the above claim of the type of words affecting readability, we calculate the mltd_ma_wrap_aw only for function words. We calculate the number of function words required to reach the required TTR value. As per the results obtained, the mltd_ma_wrap_aw in

Table 4 Number of Epochs required for different training algorithms

Dependent Variables	SCG	LM	GD
Readability	4312	2112	2425
Trustworthiness	3401	3398	2500
Neutrality	2209	4898	2678
Completeness	1924	2314	2421

[4] Content words are those words which do not have any meaning in general.

[5] Function words are those words that have a specific meaning

Table 5 Training and Testing Accuracies with different training algorithms

Sr. No.	Dependent Variables	SCG		LM		GD	
		Training	Testing	Training	Testing	Training	Testing
1	Readability	85.1	84.2	84.5	82.1	85.4	84.7
2	Trustworthiness	89.0	88.2	86.1	83.4	88.2	87.2
3	Neutrality	88.2	87.2	81.2	79.2	85.6	83.2
4	Completeness	78.1	76.2	86.8	86.2	88.6	86.3

the case of function words is also positively correlated with readability values (correlation coefficient=0.63). This again proves the relationship between lexical diversity (even in the case of function words) and readability values.

5.2 Experiment 2: mediation analysis of NLP-based parameters

As per the above discussion, it is clear that lexical diversity affects the readability of the underlying text the most. Thus, we can alter the lexical diversity to achieve the desired readability level of the underlying text. However, it should be noted that there are a number of crowdsourced parameters in Wikipedia articles that are responsible for the quality of the text and, thus, the readability of the text. We, therefore, propose a methodology that takes into account both crowdsourced parameters and lexical diversity and their effect on readability.

There have been many instances of crowdsourcing where a group of individuals outperform a single individual. One of the most famous examples is the DARPA balloon challenge, where various people across the US were able to identify the locations of balloons in different parts of the country. Accord-

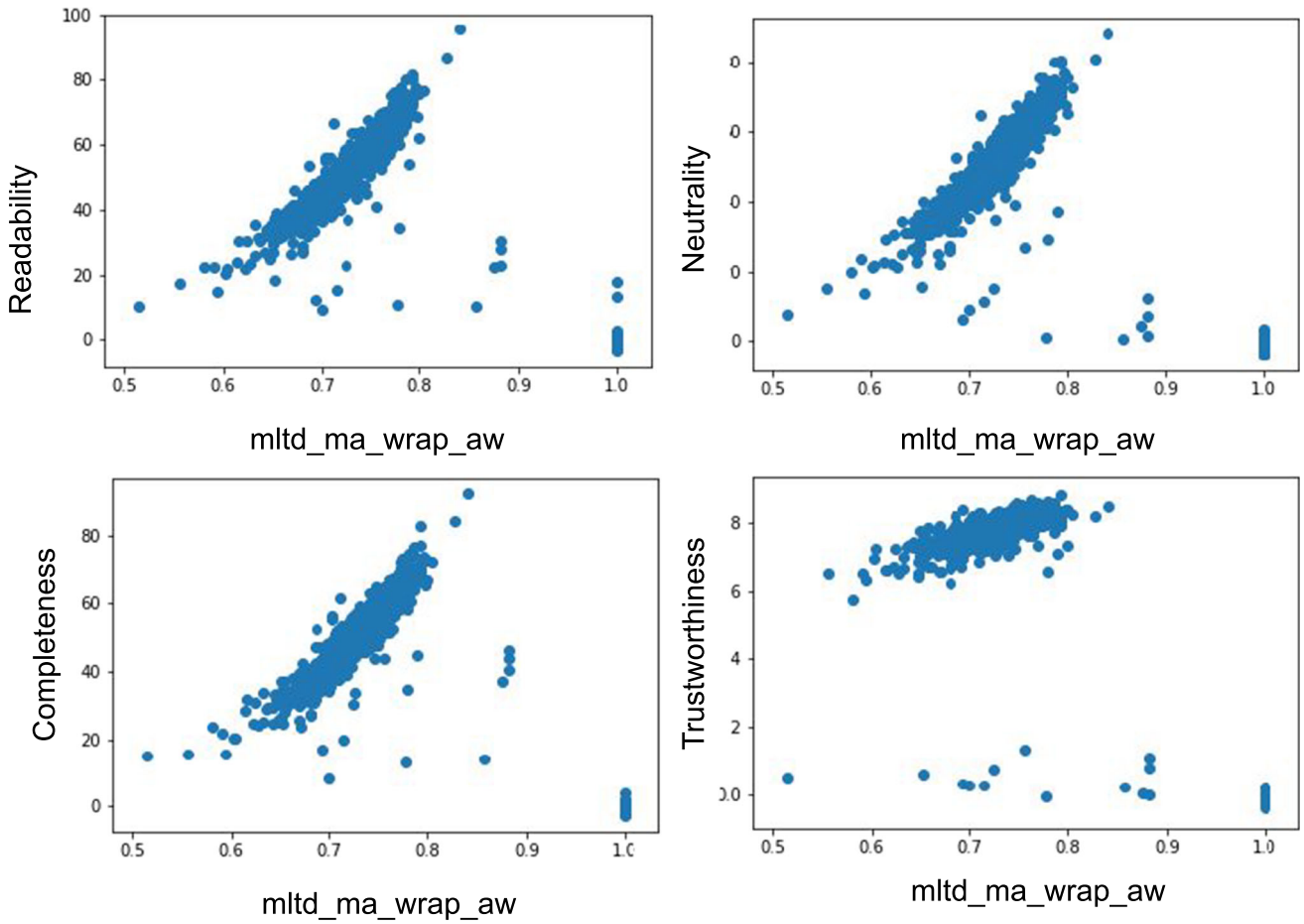


Fig. 4 The plots show the correlation of the target variable with the most dominant feature, i.e., mtld_ma_wrap_aw, that reflects the lexical diversity. The above figure reports a positive correlation between mtld_ma_wrap_aw and each of the rating dimensions

ing to 'Wisdom of Crowds' [67], it is the diversity among the group of individuals that makes them more intelligent than a single individual. Similarly, in crowdsourced platforms like Wikipedia, diversity among article editors leads to good-quality Wikipedia articles. Since quality is directly related to the readability of the articles, diversity does influence the readability of Wikipedia articles. As per the past studies, the diversity in Wikipedia articles can be measured using a number of parameters, which are described below:

- As described by Kittur and Kraut [68], implicit coordination among editors helps in the introduction of information diversity in the underlying article. The implicit coordination is quantified using the Gini coefficient of contribution done by the editors. A very similar metric called Local Workload Diversity was used by Robert et al. in one of the past studies.
- Another study focused on the quality of contribution made by the editors rather than a number of edits [69]. The experience of the editors is taken into account while judging the quality of edits made by the editor. The experience is, in turn, measured using the variation in edits made by the editor across all Wikipedia articles.
- Ren et al. coined tenure disparity as one of the measures of diversity in Wikipedia [50]. Tenure disparity quantifies the amount of time spent by the editor in editing Wikipedia articles.

To summarize, the aforementioned parameters, such as tenure disparity, implicit coordination, and experience diversity, play an important role in the quality of the articles. It should be noted that all of the above parameters are, in a way, related to the number and quality of edits. Hence, based on the previous studies, we majorly take into account two crowdsourced parameters that can capture the diversity in Wikipedia articles, i.e., the number of edits and the quality of edits. The number of edits is captured by the standard deviation of edits done by all editors. The standard deviation helps to measure the variation in edits made by the editors across all the Wikipedia articles. For an experienced editor, the standard deviation of edits must be low. The quality of edits is captured by the ratio of experienced editors editing the particular Wikipedia article. A higher number of experienced editors indicates quality edits being made to the underlying articles. Therefore, we introduce two crowdsourced parameters, one concerning the number of edits and the other one concerning the quality of edits contributed by the editors to a Wikipedia article. The details of the crowdsourced parameters are as follows:

- Standard Deviation of edits contributed by the editors of a Wikipedia article: The contribution of different editors is not uniform in a Wikipedia article. It is established by the previous literature that there are a handful of editors contributing most of the content to the article [70]. A majority of the editors are responsible for minor changes, and a small set of editors is responsible for the majority of the changes in the underlying article. However, the number of editors contributing the majority of the content varies from article to article.
- Ratio of experienced editors contributing to a Wikipedia article: The second crowdsourced parameter takes into account the experience of editors contributing to a Wikipedia article. The editors from the list of top 10,000 editors ordered by the number of edits are considered to be experienced [71]. The ratio of experienced editors is calculated as the ratio of the number of experienced editors to the total number of editors present in the Wikipedia article.

Based on the above discussion, we propose the following four hypotheses discussing the relationship between the above-crowdsourced parameters and the readability of Wikipedia articles.

H1: The standard deviation of edits contributed by the editors is positively correlated with the readability of Wikipedia articles In [72], a classification model is trained to determine the effect of writing style on the quality of the article. The results showed that the writing style of an article is an effective predictor of its quality, and as quality is directly correlated with readability, it also affects readability. [4] states that different editors have different writing styles, and therefore, a change in the writing style of an article is observed as it is crowdsourced. The changes in writing style lead to changes in the readability levels of the article.

If we test hypothesis H1, we observe a positive correlation between the standard deviation of edits contributed by the editors and the readability of articles (Correlation coefficient= 0.45). The reason is frequent change in writing style observed for articles with low standard deviation of edits which leads to low readability. If the standard deviation of the edits contributed by the editors is less, then it means all the editors are contributing uniformly to the article. Since all the editors have uniform contributions, there are frequent changes in writing style, which affects the readability levels. On the contrary, if the standard deviation of edits done by editors is high, then it means editors are contributing to the articles in a non-uniform fashion. Some editors are contributing more than the other editors. In such cases, some of the past studies claim that most of the content is added by a handful of editors, and the rest of the editors are involved in rectifying minor errors in the text [73]. The reader experiences fewer changes in the writing style, which makes the comprehension of the underlying text easier. A sudden change in the writing style while reading the underlying article hinders the comprehension of the text and affects its readability. Our claim

is also substantiated by past studies which state writing style affects readability and Wikipedia articles experience changes in the writing style [73]. Hence, a high standard deviation in edits done on the article by various editors leads to high readability in Wikipedia articles.

H2: The ratio of experienced editors is negatively correlated with the readability of Wikipedia articles

It is clear that experienced editors contribute significantly to the articles. If the ratio of experienced editors is higher in an article, then more editors are contributing heavily to the articles. Due to a higher number of editors contributing heavily to the articles, there are more frequent changes in the writing style of the article. As discussed above, sudden changes in the writing style affect the readability of the Wikipedia article. This hypothesis is also supported by the negative correlation observed between the above two parameters with a correlation coefficient of around -0.56.

After defining the nature of relationships between the two crowdsourced parameters and the readability of the text present in Wikipedia articles, we propose the following hypotheses defining the relationship between lexical diversity and the crowdsourced parameters.

H3: The standard deviation of edits contributed by the editors is negatively correlated with the Lexical Diversity of Wikipedia articles

A low standard deviation value means that all editors contribute evenly to the articles. In simple words, such an article comprises text written by various editors where the contribution made by all the editors is nearly the same. As each one of these editors uses a different vocabulary, the underlying article exhibits varied vocabulary that leads to an increase in the lexical diversity of Wikipedia articles. The above claim is also supported by the Wikipedia articles with a correlation coefficient of around -0.67.

H4: The ratio of experienced editors is positively correlated with the lexical diversity of the Wikipedia articles

A higher ratio of experienced editors in Wikipedia articles suggests that more editors contributed heavily to the underlying articles. Since all of the editors possess different vocabulary, the underlying text in the concerned Wikipedia article experiences higher lexical diversity. Also, if we calculate the correlation between the ratio of experienced editors and the lexical diversity of the article, it is around 0.48.

5.2.1 Details of Mediation Analysis

According to the neural network proposed in the last section, lexical diversity is a deciding parameter for the readability level of the text. However, it is also important to note that it is the editors who are responsible for introducing lexical diversity in the text present in Wikipedia articles. Thus, we propose a theoretical model where the various crowdsourced parameters act as independent variables, lexical diversity acts as a mediator, and the readability of the text acts as a dependent variable. The crowdsourced parameters considered are the standard deviation of edits contributed by the editors and the ratio of experienced editors editing a Wikipedia article. After testing the correlation between independent variables and the mediator variable, we perform bootstrap mediation analysis to study the mediating effects of lexical diversity on the readability of Wikipedia articles [74]. Along with the above-mentioned variables, we also introduce control variables, such as the age and size of the article, to the bootstrap mediation model. The age of the article is calculated as the time difference (in days) between the first edit and the last edit of the article. Article size is calculated as the number of words added to the article (log base-10 transformed). We use the Pingouin library in Python to perform the bootstrap mediation analysis [75]. The library uses the method proposed by Baron and Kenny for the mediation analysis. In addition to this, it should also be noted that the VIF values obtained for the crowdsourced parameters are below 10, suggesting no issues of multicollinearity in the proposed model.

In addition to the above model, we regress the two crowdsourced parameters to predict the readability values without considering the mediating effects of lexical diversity. The regression model also proves the above-stated hypotheses (H1, H2). We also regress the two crowdsourced parameters to predict the lexical diversity of the underlying text present in Wikipedia articles. This regression model also proves the above-stated hypotheses (H3, H4). The results obtained for both the models are discussed in the next section.

5.2.2 Results of experiment 2

To test hypotheses H1 and H2, a simple regression analysis is performed with the crowdsourced parameters as the independent variables and the readability of the text as the dependent variable. We also add control variables to the model, i.e., Age and Size of the article. The R^2 value of the regression model is observed to be 0.65. The results resonated with the proposed hypotheses. The model predicted a positive relationship between the standard deviation of edits contributed by the editors and the readability values of the text present in Wikipedia articles. The regression results are mentioned in the equation 1 below. Readability $= i_1 + 0.92*$ Standard deviation of edits $-1.71*$ Ratio of experienced editors $-0.41*$ Size of the article $+1.1*$ Age of the article $+e_1$

To test H3 and H4, we perform regression analysis with the standard deviation of edits contributed by the editors and ratio of experienced editors as the independent variables and lexical diversity as the dependent variable. The results show a negative relationship between lexical diversity and standard deviation of edits contributed by the editors, as hypothesized in H3. In addition, the results prove a positive relationship between lexical diversity and the ratio of experienced editors,

similar to hypothesis H4. The results are shown in Equation 2 below.

Lexical Diversity $= i_2 - 0.6*$ Standard deviation of edits $+0.51*$ Ratio of experienced editors $-0.30*$ Size of the article $+0.04*$ Age of the article $+e_2$

In addition to the above models, we also perform bootstrap mediation analysis with all the variables as depicted in Fig. 5. The bootstrap mediation analysis takes into account the mediating effects of lexical diversity on the readability of the text in Wikipedia articles. As the effects of the standard deviation of edits and the ratio of experienced editors are still significant after adding the mediator lexical diversity, we conclude that lexical diversity has a partial mediation effect on the readability of the text present in Wikipedia articles. The results for this model are shown in the equation 3 below.

Readability$= i_3 - 0.73*$ Standard deviation of edits$-1.25*$ Ratio of experienced editors $+0.62*$ Lexical Diversity $-0.35*$ Size of the article $+0.92*$ Age of the article $+e_3$

To summarise, the above results conclude the effect of crowdsourced parameters on the readability scores. As per equation 1, a higher value of the ratio of experienced editors leads to low readability scores, and a higher value of standard deviation of edits leads to high readability scores. As per equation 2, a higher value of standard deviation of edits leads to low lexical diversity, and a higher ratio of experienced editors leads to high lexical diversity. Now, as per equation 3, we still observe a significant effect of crowdsourced parameters on readability even after taking lexical diversity into account. We also observe a moderate effect of lexical diversity on readability scores. Hence, we conclude that the effect of crowdsourced parameters on readability scores is partially mediated by lexical diversity. Since there exists a partial mediation effect of lexical diversity, we can say that the effect of crowdsourced parameters on readability scores is not only because of changes observed in the writing style of the article. There are other factors which are responsible for the effect of crowdsourced parameters on readability scores. One of the probable reasons behind the effect of crowdsourced parameters on readability scores can be conflicts among the editors. In the case of a low standard deviation of edits and a high ratio of experienced editors, we observe an equal number of edits by all the editors. In such a scenario where editors have an equal number of edits, there is collective ownership of articles, which often leads to conflicts among editors. With a high number of conflicts among editors, much of the time/attention is wasted on conflict resolution, and article quality suffers. Hence, we experience low readability in such articles. It should be noted that the above explanation does not consider the effect of writing style or lexical diversity. It takes into account the effect of crowd dynamics on the readability scores. Hence, apart from crowdsourced parameters affecting the lexical diversity and lexical diversity affecting readability scores, the crowdsourced parameters also directly influence the readability scores.

6 Conclusion

The regression models built with NLP parameters as the input features and trustworthiness, neutrality, readability, and completeness scores as the target variables show that the NLP-based features do affect the readability score of a crowdsourced document. The NLP-based features are calculated using a wide range of tools, such as TAALED, TAASSC, TAALES, SEANCE, and TAACO. Further, the above exercise establishes lexical diversity as one of the deciding factors behind a good readability score. The past studies also show that crowdsourced parameters influence the readability scores of Wikipedia articles. Hence, it is clear that both NLP-based and crowdsourced parameters affect the readability scores of the articles. However, it should be noted that it is the crowd only that is responsible for the generation of the entire article in Wikipedia, and the effect of crowdsourced parameters on NLP-based parameters can not be ignored either. We conduct bootstrap mediation analysis to judge the effect of crowdsourced parameters on NLP-based parameters, which in turn influence the readability scores. The results show that the NLP-based features partially

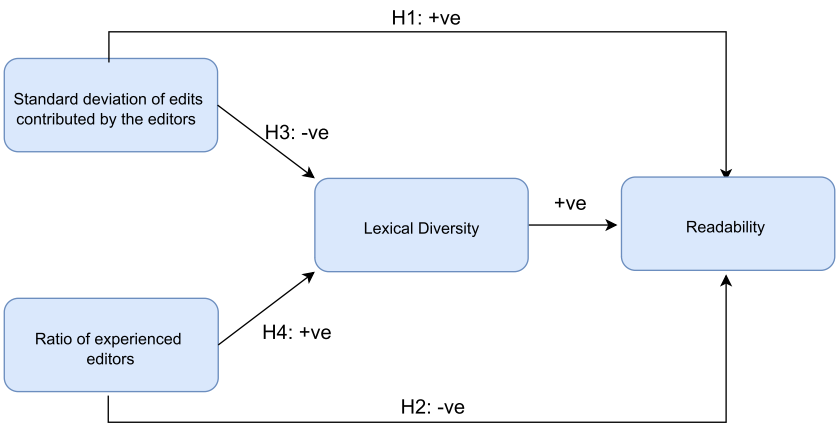


Fig. 5 Bootstrap Mediation Analysis with the readability as the dependent variable, lexical diversity as the mediator variable, and standard deviation of edits and the ratio of experienced editors as the independent variables

mediate the effect of crowdsourced parameters on readability scores. Therefore, it is important to consider the effect of both crowdsourced and NLP-based parameters while calculating the readability scores of Wikipedia articles.

In the future, we plan to study the changes in lexical diversity with every change introduced in the crowdsourced article and their effect on readability scores. Additionally, different crowdsourced parameters, such as conflicts observed between editors and interaction among the editors, can also be considered to quantify their effect on readability scores. It will help us better understand the crowd dynamics that affect the NLP parameters, leading to a change in the readability scores.

Appendix A: tools and techniques

As discussed earlier, conventional readability metrics consider only surface-level parameters to calculate the text readability. In order to calculate non-trivial parameters related to the readability of the text, like lexical diversity, lexical complexity, and semantic complexity (as discussed in the section Related Work), we use a number of text analysis tools. The tools used are TAALED, SEANCE, TAALES, TAASSC, and TAACO, which are discussed below.

1. TAALED (Tool for the Automatic Analysis of Lexical Diversity): TAALED is used to quantify different measures related to lexical diversity [52][6]. For this, the author defines three dimensions of lexical diversity: volume, abundance, and variety. These dimensions are calculated as follows:

   - Volume: Volume is defined as the number of tokens present in the text. It is calculated using the following expression.

   $$Volume = \sum_{i=1}^{L_w} w_i \; ; \; 1 < i < L_w$$

   where w_i refers to the i^{th} word in the text and Lw refers to the length of the document in words.
   - Abundance: Abundance refers to the total number of different types that exist in the text. The types of words are discerned using lemmatization. Lemmatization refers to converting the underlying text to various tokens and finding the root word of these tokens. For example, the root word of sleeping and slept is sleep, and both of these words are considered under the same type, i.e., sleep. In short, the lemmas (or words discerned through lemmatization) refer to different types of words present in the text. Abundance for a given text is calculated as follows:

   $$Abundance = \sum_{i=1}^{L_l} t_i$$

   $$t_j = \begin{array}{r} t_j + 1 \; ; \; w_i \in t_j \; ; \; 1 < i < L_w \; ; \; 1 < j < L_l \\ t_j \quad ; \; w_i \notin t_j \; ; \; 1 < i < L_w \; ; \; 1 < j < L_l \end{array}$$

   where t_j refers to the j^{th} type/lemma in the text, and Ll refers to the total number of lemmas present in the text.
   - Variety: For measuring the variety, the authors define a number of indices. The indices are HD-D, MATTR, and MTLD.

Algorithm 1 Algorithm to compute MTLD.

```
procedure COMPUTE MTLD(L_w, Type, Token)
    L_w : Length of document in words
    Type : List of types of words in document
    Token : List of tokens in document
    i = 0
    TTR_0 = Type[0] / Token[0]
    for i < L_w do
        if TTR_i = TTR_{i+1} then
            break
        else
            TTR_i = (sum_{j=1}^{i} Type[J]) / (sum_{j=1}^{i} Token[J])
        end if
    end for
end procedure
```

Table 6 Indices for Syntactic Complexity Analyzer

Measure	Definition
Mean Length of Clause	$\frac{No.of words}{No.of clauses}$
Mean Length of Sentence	$\frac{No.of words}{No.of sentences}$
Mean Length of T-units	$\frac{No.of words}{No.of T-units}$
Sentence Complexity Ratio	$\frac{No.of clauses}{No.of sentences}$
T-unit Complexity Ratio	$\frac{No.of clauses}{No.of T-units}$
Dependent Clause Ratio	$\frac{No.of dependent clauses}{No.of clauses}$
Dependent Clauses per T-unit	$\frac{No.of dependent clauses}{No.of T-units}$
Coordinate phrases per clause	$\frac{No.of coordinate phrases}{No.of clauses}$
Coordinate phrases per T-unit	$\frac{No.of coordinate phrases}{No.of T-units}$
Sentence Coordination Ratio	$\frac{No.of T-units}{No.of sentences}$
Complex Nominals per clause	$\frac{No.of complex nominals}{No.of clauses}$
Complex Nominals per T-unit	$\frac{No.of complex nominals}{No.of T-units}$

[6] https://github.com/kristopherkyle/TAALED

– HD-D refers to the probability of a word being included in a random sample from the text. The probability is calculated using the hypergeometric distribution. The hypergeometric distribution computes the probability of a word occurring i times out of n trials when a sample is drawn from a population without replacement. The probability of a word occurring i times out of n trials is calculated as follows:

$$P(i|N,m,n) = \frac{\binom{m}{i}\binom{N-m}{m-i}}{\binom{N}{n}}$$

where N is the size of the population, n is the number of items in the sample, i is the number of w_i words in the sample, m is the number of successes, i.e., the number of samples comprising of w_i word.

– MATTR: MATTR (Moving Average Type Token Ratio) calculates the moving average in the context of the type-token ratio. It calculates TTR by averaging over equally sized multiple overlapping windows. As stated by the previous studies, MATTR is calculated over a window size of 50. This leads us to the following expression used to calculate MATTR.

$$MATTR = \frac{\frac{\sum_{j=1}^{L_t} t_j}{\sum_{i=1}^{50} w_i}}{50}$$

where L_t refers to the total number of types present in each of the windows with size 50.

– MTLD: MTLD measures the number of words required to reach a certain value of TTR (Type Token Ratio). According to previous studies, the optimal TTR value is 0.720 [52]. Algorithm 1 is used to calculate MTLD.

2. TAASSC (Tool for the Automatic Analysis of Syntactic Sophistication and Complexity): TAASSC is a syntactic analysis tool that measures 367 indices related to syntactic sophistication and complexity[7]. These indices can be divided into four categories:

   - Syntactic Complexity Analyzer (SCA) indices: In 2010, Lu [76] developed a tool to measure the syntactic complexity called Syntactic Complexity Analyzer. The author introduced a number of indices counting different syntactic structures which can be used to measure syntactic complexity. These indices are summarised in Table 6. The indices given in the table are calculated, which correlate with the syntactic complexity of the underlying text. The definitions of terminologies used to calculate syntactic complexity are as follows.

     – Clause: A part of the sentence comprising of subject and verb. For example, in the sentence 'I am eating ice-cream', the clause is 'I am eating.'
     – T-units: The T-units refers to the shortest text that the given sentence can be split into according to grammatical rules. For example, in the sentence 'He is the head of the CSE department', 'He is the head' can be regarded as a T-unit.
     – Dependent Clauses: They are a special type of clauses that do not form a sentence according to grammatical rules. For example, in the sentence 'I went on a car that my husband gifted me.', the clause 'that my husband gifted me' is a dependent clause.
     – Coordinate Phrases: The phrases that are joined by coordinating conjunctions, such as for, and, but, nor, or, yet.
     – Complex Nominals: Complex nominals are normally a combination of nouns and adjectives. For example, beautiful house, wind turbine, and many more such phrases.

   - Fine-grained indices of clausal complexity: In addition to the indices defined by SCA, TAASSC also defines a number of indices that help measure the clausal complexity in the text. It should be noted that clausal complexity in TAASSC is measured in terms of the number of dependent types present in the clause in contrast to SCA, where the clausal complexity is measured as the number of words in the clause [77]. This prevents the clauses with more words from being given higher weight as compared to the clauses with fewer words. The various dependent types defined are adjectives (for example, in the sentence "She looks beautiful" , beautiful is an adjective dependent type.), adverbs (for example, in the sentence "Accordingly, I decided to eat burger", accordingly is an adverb dependent type.) auxiliary verb (for example, in the sentence "She was running", was is an auxiliary dependent type.).
   - Fine-grained indices of phrasal complexity: TAASSC includes a number of indices measuring phrasal complexity dependent on noun phrases and dependent types [77]. Some of the noun phrases considered are nominal subject, prepositional object, passive nominal object, and so on. An example of a nominal subject is "The man in a red shirt gave money to the beggar". Here 'the man in the red shirt' is a nominal subject. Three types of indices are calculated using the above noun phrases and dependent types.

[7] https://github.com/kristopherkyle/TAASSC

(a) The first type calculates the average number of dependents per each noun phrase (i.e., the nominal objects).
(b) The second type calculates the occurrence of particular dependent types (i.e., adjectives) present in the text.
(c) The third phrasal index type calculates the average occurrence of particular dependent types in particular noun phrases (i.e., adjectives occurring in nominal objects).

- Frequency-based indices of syntactic sophistication: Syntactic sophistication is defined as the relative difficulty encountered in comprehending the text [78]. TAASSC calculates 15 basic indices related to syntactic sophistication [77]. Each of the indices involves some variation, resulting in 38 indices. Also, each of these indices is calculated with respect to 5 different corpora (written, academic, fiction, magazine, and newspaper). Hence, the total number of indices calculated is 190. Verb lemmas and Verb Argument Constructions (VAC) are used to quantify syntactic sophistication. VACs define the structure of the sentences. For example, in the sentence 'He kicked the ball', VAC is subject-verb-object. The measures defined in TAASSC are related to frequency, conditional probability of verb, and VAC occurring together, and many more.

3. TAALES (Tool for the Automatic Analysis of Lexical Sophistication): Lexical Sophistication takes into account the length and breadth of lexical words so that it does not only consider the frequency of lexical words but also a number of quality indices, such as familiarity, meaningfulness, imageability [53][8]. The indices used in TAALES are defined below.

   - Word Frequency: Word frequency measures the number of times a particular word occurs in a corpus of texts. The rationale behind using word frequency as one of the measures for lexical sophistication is that words that are less frequent are considered to be more sophisticated than the words that occur frequently. For example, the commonly occurring words, like give, take, and create, are considered less sophisticated as compared to the other words, which occur less frequently.
   - Word Range: Word Range is calculated as the number of texts present in the corpus in which a particular word occurs. Word Range (WR_i) for the i^{th} word (w_i) calculated as follows:

$$WR_i = \begin{matrix} WR_i + 1 \; ; \; w_i \in D_i \\ WR_i \quad ; \; w_i \notin D_i \end{matrix}$$

     where D_i refers to i^{th} text in the corpus.
   - N-gram Frequency: The n-gram frequency measures the multi-word units present in the text. The length of multi-word units is termed as n. For example, in the sentence 'I had a great day', the 3-gram units are 'I had a', 'had a great', 'a great day'. The N-gram frequency measures the number of n-gram units for different values of n.
   - Academic Language: In academic language, there are certain words that occur more frequently as compared to the words frequently occurring in general language. For example, a technical document on *Stem Cells* would comprise various technical terms related to Biology that are not used in general language. Academic Word List (AWL) is a list of such words that maintains a record of frequently occurring words in academic language. A higher proportion of such words in the text leads to a more sophisticated text.
   - Psycholinguistic Word Formation: The psycholinguistic word formation helps to model the cognitive aspects of a word. These cognitive aspects help to measure the quality of writing and lexical sophistication of the underlying text. The various cognitive aspects captured by psycholinguistic word formation are as follows.

     – Concreteness: A word is said to be concrete if it refers to a person or an object. The abstract words, such as love, anger, and hate, score low on concreteness.
     – Familiarity: The frequently occurring words (such as give, take, help) in the English language are considered to be more familiar than the ones that are less frequent.
     – Imagability: Those words are said to be imagable, which produce an image in mind immediately, like a cow, car, or mountains.

4. TAACO (Tool for the Automatic Analysis of Cohesion): TAACO computes a number of indices that help measure text cohesion [54][9]. Text cohesion is critical to the readability of the underlying text as it allows the reader to make connections between paragraphs/sentences. These connections help the reader to

[8] https://www.linguisticanalysistools.org/taales.html

[9] https://www.linguisticanalysistools.org/taaco.html

infer that similar concepts/ideas are conveyed across different paragraphs/sentences. These external cues can either be lemma overlap or semantic overlap. Based on these two types of overlap, cohesion indices are defined as follows.

- Lexical Overlap: The lexical overlap is measured using overlapping lemmas between adjacent sentences and paragraphs. The overlap scores are calculated by quantifying the intersecting POS tags, such as nouns, verbs, and adverbs, between adjacent paragraphs/sentences. Algorithm 2 is used to calculate lexical overlap [54].

Algorithm 2 Algorithm to compute Lexical Overlap.

procedure LEXICAL OVERLAP(Sentences_List)
  TL : *List of Tokens in each of the sentences.*
  Lw : *Length of the document in words.*
  TC : *Dictionary of different token type count*
  LO : *Dictionary of lexical overlap scores*
  $i = 0$
  for $i < L_w$ **do**
    $TL[i]$=$Sentences\_List[i].tokenize()$
  end for
  for $i < L_w$ **do**
    if $TL[i] \bigcap TL[i+1] \neq \emptyset$ **then**
      IT=$TL[i] \bigcap TL[i+1]$
    end if
    for $j < length(IT)$ **do**
      $TC[type(IT[i])]$=$TC[type(IT[i]] + 1$
      $LO[type(IT[i])]$=$\frac{TC[type(IT[i])]}{L_w}$
    end for
  end for
end procedure

- Semantic Overlap: Apart from lexical overlap, TAACO also calculates semantic overlap. Semantic overlap is also calculated across sentences and paragraphs, but the overlap is considered in case semantically similar words/lemmas are encountered across sentences and paragraphs. Semantically similar words are figured out using WordNet database [79]. For calculating the semantic overlap, the intersection between subsequent sentences/paragraphs is calculated by considering semantically similar words between subsequent sentences/paragraphs. Algorithm 3 is used to calculate semantic overlap [54].
- Givenness: In addition to this, the givenness of underlying text is also calculated, which measures the amount of information that can be recovered from the previous discourse. The frequency of a variety of pronoun types is measured to quantify givenness under the assumption that pronouns are used when information is conveyed in the text [10].

Algorithm 3 Algorithm to compute Semantic Overlap.

procedure SEMANTIC OVERLAP(Sentences_List)
  TL : *List of Tokens in each of the sentences.*
  L_w : *Length of the document in words.*
  TC : *Dictionary of count of different types of tokens.*
  SO : *Dictionary of semantic overlap scores*
  Syn : *Dictionary of synonyms for each token in the document*
  $i = 0$
  for $i < L_w$ **do**
    $TL[i]$=$Sentences\_List[i].tokenize()$
  end for
  for $i < L_w$ **do**
    for $eachinTL[i]$ **do**
      $Syn\_Tokens = SYN[each]$
      if $Syn\_Tokens \bigcap TL[i+1] \neq \emptyset$ **then**
        IT=$Syn\_Tokens \bigcap TL[i+1]$
      end if
    end for
    for $j < length(\texttt{IT})$ **do**
      $TC[type(IT[i])]$=$TC[type(IT[i]] + 1$
      $LO[type(IT[i])]$=$\frac{TC[type(IT[i])]}{L_w}$
    end for
  end for
end procedure

- Connectiveness: It is used to measure the frequency of connective indices and the connective indices in the underlying text use the indices present in Coh-Metrix [80], such as minimal edit distance, readability scores such as Flesch-Kincaid score, concreteness of content words, and similar other indices.

5. SEANCE (Sentiment Analysis and Social Cognition Engine): Sentiment analysis refers to the extraction of information related to human feelings, emotions, and opinions from the text [81][10]. SEANCE contains a number of pre-developed vectors that are used to measure the sentiment and social cognition of the given text. The vectors are extracted from freely available databases such as SenticNet [55, 56], VADER [57], Hi-Liu polarity [58], Emolex [59]. A number of pre-trained models like BERT are also used to extract the sentiment from the text [31, 82]. SEANCE calculates 3000 indices related to sentiment analysis, which are further categorized into 80 categories [21]. However, due to overlap among the aforementioned databases, some redundant indices are also present among 3000 indices. The dimensionality reduction algorithm, known as Principal Component Analysis (PCA), is applied to reduce the number of indices. PCA helps reduce a large number of variables to a set of 40 derived variables. The other two features of SEANCE are as follows.

[10] https://github.com/kristopherkyle/SEANCE

- Unlike the other sentiment analysis tools like LIWC, SEANCE also takes into consideration the negation factors. Sometimes, the sentences convey negative sentiment with a positive word by using negations. For example, The movie was not good. Here, good is a positive word but is considered negative due to the negation word. So, SEANCE considers round three words preceding a particular word to capture the negation words, which may change the sentiment of the underlying text.
- SEANCE also includes Stanford POS tagger. POS is an important component of sentiment analysis because unique aspects of sentiment may be conveyed more strongly by adjectives, verbs, or adverbs.

The above-mentioned tools are used to calculate the indices for measuring lexical diversity, semantic complexity, lexical sophistication, cohesion, and sentiments. As discussed in Section 2, the aforementioned NLP parameters affect the readability of the text. We train a neural network with these NLP parameters to determine the relative effect of all these features on the readability scores. Next, we discuss the dataset and the experiments carried out as a part of this study.

Appendix B: Permutation importance in ANN

This section describes the relative importance of features used in ANN. The tables below (Table 7 and Table 8) enumerate the ten most and least important features in ANN.

Table 7 Least important features in ANN

Index	Relative Drop in Accuracy
TL_Freq_AW	0.001
KF_Nsamp_AW	0.002
TL_Freq_AW_Log	0.041
Brown_Freq_AW_Log	0.072
adjacent_overlap_cw_sent	0.218
MRC_Concreteness_CW	0.231
adjacent_overlap_all_sent	0.250
adjacent_overlap_cw_sent	0.218
MRC_Concreteness_CW	0.231
Brown_Freq_FW_Lo	0.238

Table 8 Most important features in ANN

Index	Relative Drop in Accuracy
mtld_ma_wrap_aw	26.127
mtld_ma_bi_aw	25.021
mattr50_aw	17.215
maas_ttr_aw	8.758
log_ttr_aw	3.071
positive_nouns_component	2.850
politeness_component	2.843
Skltot_Lasswell	2.755
basic_ntokens	2.202
basic_ncontent_types	2.131

Author Contributions Simran Setia: Conceptualization, Methodology, Software, Investigation, Writing - Original Draft. Anamika Chhabra: Conceptualization, Methodology. Amit Arjun Verma: Conceptualization, Methodology. Akrati Saxena: Methodology, Writing - Review & Editing, Supervision.

Data Availability and Access This work has used data from the following source: https://www.mediawiki.org/wiki/Article_feedback/Version_4, and it is publicly available. The Wikipedia articles used as a part of the dataset are also available online on Wikipedia (https://en.wikipedia.org/).

Declarations

Ethical and informed consent for data used This study does not contain any studies with human or animal subjects performed by any of the authors.

Competing Interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. interest.

References

1. Zuffi S, Brambilla C, Beretta G, Scala P (2007) Human computer interaction: Legibility and contrast. In: 14th International conference on image analysis and processing (ICIAP 2007), IEEE, pp 241–e246
2. Alexa (2019) Wikipedia.org Traffic, Demographics and Competitors. https://www.alexa.com/siteinfo/wikipedia.org

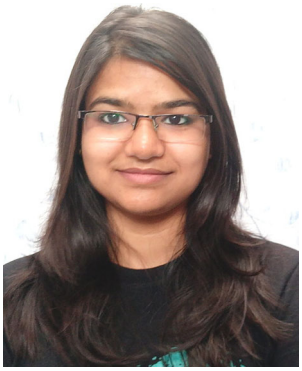
3. Swartz A (2006) Who writes wikipedia. Raw thought 4
4. Setia S, Iyengar S, Verma AA, Dubey N (2021) Is wikipedia easy to understand?: A study beyond conventional readability metrics. In: International Conference on Computational Collective Intelligence, Springer, pp 175–e187
5. Gregori-Signes C, Clavel-Arroitia B (2015) Analysing lexical density and lexical diversity in university students' written discourse. Procedia Soc Behav Sci 198:546-e556
6. Todirascu A, François T, Gala N, Fairon C, Ligozat A-L, Bernhard D (2013) Coherence and cohesion for the assessment of text readability. In: Proceedings of 10th international workshop on natural language processing and cognitive science (NLPCS 2013), pp 11–e19
7. Rezaee AA, Norouzi MH (2011) Readability formulas and cohesive markers in reading comprehension. Theory & Practice in Language Studies 1(8)
8. Zhang H, Gan W, Jiang B (2014) Machine learning and lexicon based methods for sentiment classification: A survey. In: 2014 11th Web Information System and Application Conference, IEEE, pp 262–e265
9. Crossley SA, Greenfield J, McNamara DS (2008) Assessing text readability using cognitively based indices. TESOL Q 42(3):475-e493
10. Crossley S, McNamara D (2014) Developing component scores from natural language processing tools to assess human ratings of essay quality. In: The twenty-seventh international flairs conference
11. Cobb T (2007) Computing the vocabulary demands of l2 reading. Language Learning & Technology 11(3):38-e63
12. Eslami H (2014) The effect of syntactic simplicity and complexity on the readability of the text. J Lang Teach & Res 5(5)
13. McNamara DS, Kintsch W (1996) Learning from texts: Effects of prior knowledge and text coherence. Discourse Process 22(3):247-e288
14. McNamara D, Kintsch E, Songer N, Kintsch W (1996) Are good texts always better? text coherence, background knowledge, and levels of understanding in learning from text. Cogn Instr 14(1):43
15. McNamara DS, Louwerse MM, Graesser AC (2002) Coh-metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension. Technical report, Technical report, Institute for Intelligent Systems, University of Memphis
16. Britton BK, Van Dusen L, Gülgöz S, Glynn SM, Sharp L (1991) Accuracy of learnability judgments for instructional texts. J Educ Psychol 83(1):43
17. Yan X, Song D, Li X (2006) Concept-based document readability in domain specific information retrieval. In: Proceedings of the 15th ACM international conference on information and knowledge management, pp 540–e549
18. Leacock C, Chodorow M (1998) Combining local context and wordnet similarity for word sense identification. WordNet: An electronic lexical database 49(2), 265–e283
19. Wu C, Cao L, Chen J, Wang Y, Su J (2023) Modeling different effects of user and product attributes on review sentiment classification. Appl Intell, pp 1–16
20. Saxena A, Reddy H, Saxena P (2022) Recent developments in sentiment analysis on social networks: techniques, datasets, and open issues. Principles of Social Networking: The New Horizon and Emerging Challenges, pp 279–306
21. Crossley SA, Kyle K, McNamara DS (2017) Sentiment analysis and social cognition engine (seance): An automatic tool for sentiment, social cognition, and social-order analysis. Behav Res Methods 49(3):803-e821
22. Shapiro AH, Sudhof M, Wilson DJ (2022) Measuring news sentiment. Journal of econometrics 228(2):221-e243
23. Bansal B, Srivastava S (2019) Hybrid attribute based sentiment classification of online reviews for consumer intelligence. Appl Intell 49(1):137-e149
24. Hoang M, Bihorac OA, Rouces J (2019) Aspect-based sentiment analysis using bert. In: Proceedings of the 22nd nordic conference on computational linguistics, pp 187–e196
25. Tan L, Tan OK, Sze CC, Goh WWB (2023) Emotional variance analysis: A new sentiment analysis feature set for artificial intelligence and machine learning applications. PLoS ONE 18(1):0274299
26. Yano Y, Long MH, Ross S (1994) The effects of simplified and elaborated texts on foreign language reading comprehension. Lang Learn 44(2):189-e219
27. Solomon RL, Howes DH (1951) Word frequency, personal values, and visual duration thresholds. Psychol Rev 58(4):256
28. Richardson JT (1975) The effect of word imageability in acquired dyslexia. Neuropsychologia 13(3):281-e288
29. Besharati MR, Izadi M (2021) Dastex: a new readability formula based on semantic complexity of text
30. Lu X (2011) A corpus-based evaluation of syntactic complexity measures as indices of college-level esl writers' language development. TESOL Q 45(1):36-e62
31. Zhang T, Gong X, Chen CP (2021) Bmt-net: Broad multitask transformer network for sentiment analysis. IEEE transactions on cybernetics 52(7):6232-e6243
32. Thierry N, Bao B-K, Ali Z, Tan Z, Christ Chatelain IB, Kefalas P (2023) Prm-kged: paper recommender model using knowledge graph embedding and deep neural network. Appl Intell pp 1–15
33. To V, Fan S, Thomas D (2013) Lexical density and readability: A case study of english textbooks. Internet Journal of Language, Culture and Society 37:61–71
34. Crossley SA, Skalicky S, Dascalu M (2019) Moving beyond classic readability formulas: New methods and new models. J Res Reading 42(3–4):541-e561
35. Kyle K, Sung H, Eguchi M, Zenker F (2023) Evaluating evidence for the reliability and validity of lexical diversity indices in l2 oral task responses. Stud Second Lang Acquis pp 1–22
36. Woods K, Hashimoto B, Brown EK (2023) A multi-measure approach for lexical diversity in writing assessments: Considerations in measurement and timing. Assess Writ 55
37. Lucassen T, Dijkstra R, Schraagen JM (2012) Readability of wikipedia. First Monday
38. Jatowt A, Tanaka K (2012) Is wikipedia too difficult? comparative analysis of readability of wikipedia, simple wikipedia and britannica. In: Proceedings of the 21st ACM international conference on information and knowledge management, pp 2607–e2610
39. Benjamin RG (2012) Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. Educ Psychol Rev 24:63-e88
40. Gkikas DC, Tzafilkou K, Theodoridis PK, Garmpis A, Gkikas MC (2022) How do text characteristics impact user engagement in social media posts: Modeling content readability, length, and hashtags number in facebook. International Journal of Information Management Data Insights 2(1):100067
41. Liang K, Liu H, Shan M, Zhao J, Li X, Zhou L (2023) Enhancing scenic recommendation and tour route personalization in tourism using ugc text mining. Appl Intell pp 1–36
42. Martinc M, Pollak S, Robnik-Šikonja M (2021) Supervised and unsupervised neural approaches to text readability. Comput Linguist 47(1):141-e179
43. Watad A, Bragazzi NL, Brigo F, Sharif K, Amital H, McGonagle D, Shoenfeld Y, Adawi M et al (2017) Readability of wikipedia pages on autoimmune disorders: systematic quantitative assessment. J Med Internet Res 19(7):8225
44. Modiri O, Guha D, Alotaibi NM, Ibrahim GM, Lipsman N, Fallah A (2018) Readability and quality of wikipedia pages on neurosurgical topics. Clin Neurol Neurosurg 166:66-e70
45. Azer SA, AlSwaidan NM, Alshwairikh LA, AlShammari JM (2015) Accuracy and readability of cardiovascular entries on

wikipedia: are they reliable learning resources for medical students? BMJ Open 5(10):008187
46. Suwannakhan A, Casanova-Martínez D, Yurasakpong L, Montriwat P, Meemon K, Limpanuparb T (2020) The quality and readability of english wikipedia anatomy articles. Anat Sci Educ 13(4):475-e487
47. Candelario DM, Vazquez V, Jackson W, Reilly T (2017) Completeness, accuracy, and readability of wikipedia as a reference for patient medication information. J Am Pharm Assoc 57(2):197-e200
48. Nassiri N, Cavalli-Sforza V, Lakhouaja A (2023) Approaches, methods, and resources for assessing the readability of arabic texts. ACM Transactions on Asian and Low-Resource Language Information Processing 22(4):1-e30
49. Jarvis S, Daller M (2013) Defining and measuring lexical diversity. Human ratings and automated measures. Amsterdam, The Netherlands, Vocabulary knowledge
50. Ren R, Yan B (2017) Crowd diversity and performance in wikipedia: The mediating effects of task conflict and communication. In: Proceedings of the 2017 CHI conference on human factors in computing systems, pp 6342–e6351
51. Gooding S, Berzak Y, Mak T, Sharifi M (2021) Predicting text readability from scrolling interactions. In: Proceedings of the conference on natural language learning
52. Kyle K, Crossley SA, Jarvis S (2021) Assessing the validity of lexical diversity indices using direct judgements. Lang Assess Q 18(2):154-e170
53. Kyle K, Crossley SA (2015) Automatically assessing lexical sophistication: Indices, tools, findings, and application. TESOL Q 49(4):757-e786
54. Crossley SA, Kyle K, McNamara DS (2016) The tool for the automatic analysis of text cohesion (taaco): Automatic assessment of local, global, and text cohesion. Behav Res Methods 48(4):1227-e1237
55. Cambria E, Speer R, Havasi C, Hussain A (2010) Senticnet: A publicly available semantic resource for opinion mining. In: 2010 AAAI Fall symposium series
56. Cambria E, Havasi C, Hussain A (2012) Senticnet 2: A semantic and affective resource for opinion mining and sentiment analysis. In: Twenty-fifth international flairs conference
57. Hutto C, Gilbert E (2014) Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the international AAAI conference on web and social media, vol 8, pp 216–e225
58. Hu M, Liu B (2004) Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, pp 168–e177
59. Mohammad S, Turney P (2010) Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In: Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text, pp 26–e34
60. Wikipedia (2013) Article Feedback. https://www.mediawiki.org/wiki/Article_feedback#Version_4
61. Wikipedia (2013) Article Feedback Tool Version 4. https://www.mediawiki.org/wiki/Article_feedback/Version_4
62. Møller MF (1993) A scaled conjugate gradient algorithm for fast supervised learning. Neural Netw 6(4):525-e533
63. Du S, Lee J, Li H, Wang L, Zhai X (2019) Gradient descent finds global minima of deep neural networks. In: International conference on machine learning, PMLR, pp 1675–e1685
64. Sapna S, Tamilarasi A, Kumar MP et al (2012) Backpropagation learning algorithm based on levenberg marquardt algorithm. Comp Sci Inform Technol (CS and IT) 2:393-e398
65. Ahmad GN, Fatima H, Ullah S, Saidi AS et al (2022) Efficient medical diagnosis of human heart diseases using machine learning techniques with and without gridsearchcv. IEEE Access 10:80151-e80173
66. Pedregosa F, Varoquaux G, Gramfort,A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al (2011) Scikit-learn: Machine learning in python. J Mach Learn Res 12:2825–e2830
67. Surowiecki J (2005) The wisdom of crowds/james surowiecki. Anchor, NY
68. Kittur A, Kraut RE (2008) Harnessing the wisdom of crowds in wikipedia: quality through coordination. In: Proceedings of the 2008 ACM conference on computer supported cooperative work, pp 37–e46
69. Yang D, Halfaker A, Kraut R, Hovy E (2016) Who did what: Editor role identification in wikipedia. In: Proceedings of the international AAAI conference on web and social media, vol 10, pp 446–e455
70. Wilkinson DM, Huberman BA (2007) Assessing the value of cooperation in wikipedia. First Monday
71. Wikipedia (2022) Wikipedia:List of Wikipedians by number of edits. https://en.wikipedia.org/wiki/Wikipedia:List_of_Wikipedians_by_number_of_edits
72. Lipka N, Stein B (2010) Identifying featured articles in wikipedia: writing style matters. In: Proceedings of the 19th international conference on world wide web, pp 1147–e1148
73. O'mahony S, Ferraro F (2007) The emergence of governance in an open source community. Acad Manag J 50(5):1079–e1106
74. MacKinnon DP, Fairchild AJ, Fritz MS (2007) Mediation analysis. Annu Rev Psychol 58:593
75. Vallat R (2018) Pingouin: statistics in python. J. Open Source Softw. 3(31):1026
76. Lu X (2010) Automatic analysis of syntactic complexity in second language writing. Int J Corpus Linguistics 15(4):474-e496
77. Kyle K (2016) Measuring syntactic development in l2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication
78. Bulte B, Housen A (2012) Defining and operationalising l2 complexity. Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA, pp 23–46
79. Miller GA (1995) Wordnet: a lexical database for english. Commun ACM 38(11):39-e41
80. Graesser AC, McNamara DS, Cai Z, Conley M, Li H, Pennebaker J (2014) Coh-metrix measures text characteristics at multiple levels of language and discourse. Elem Sch J 115(2):210-e229
81. Saxena A, Reddy H, Saxena P (2022) Introduction to sentiment analysis covering basics, tools, evaluation metrics, challenges, and applications. Principles of Soc Netw: The New Horizon and Emerging Challenges, pp 249–277
82. Gan C, Cao X, Zhu Q, Jain DK, García S (2023) Enhancing microblog sentiment analysis through multi-level feature interaction fusion with social relationship guidance. Appl Intell pp 1–17

Simran Setia received the B.Tech. degree in Computer Science and Engineering from Jaypee University of Information Technology, India in 2015. She received M.Tech. degree in Computer Science and Engineering from Thapar Institute of Engineering and Technology, India in 2017, and the Ph.D. degree in computer science and engineering from the Indian Institute of Technology, Ropar, in 2022. She works as an Assistant Professor withThapar Institute of Engineering and Technology, India. Her main areas of interest lie in Artificial Intelligence, Human-Computer Interaction, Natural Language Processing, and Crowdsourcing.

Anamika Chhabra received the B.Tech. degree in information technology from Kurukshetra University, Kurukshetra, India in 2004. She received M.Tech. degree in information technology from Guru Gobind Singh Indraprastha University, Delhi, India in 2008, and the Ph.D. degree in computer science and engineering from the Indian Institute of Technology, Ropar, in 2021. She works as a data science instructor with the Indian Institute of Technology Madras (online degree program). Her main areas of interest lie in Artificial Intelligence, Human-Computer Interaction, Social Network Analysis, and Crowdsourcing.

Amit Arjun Verma completed his PhD from the prestigious Indian Institute of Technology, Ropar, where he delved deeply into core compression problems within the domain of NLP. Throughout his doctoral research, Amit exhibited exceptional skills and innovation, leading to the development of numerous open-sourced libraries that significantly contributed to collective intelligence research. Following the successful completion of his PhD, Amit embarked on a journey in the industry, bringing his expertise to real-world applications. He joined CoreCLM, located in Seattle, as a Research Scientist, where he focused on the development of cutting-edge AI tools. Currently, Amit Arjun Verma serves as the head of the Data Science Tech team at GUVI, a role in which he continues to drive innovation and excellence.



Akrati Saxena is an assistant professor at the computer science and AI department of the Faculty of Science at Leiden University. Before joining Leiden University, she worked as a Research Fellow at NUS, Singapore and Eindhoven University of Technology, Netherlands. Her research interests include Social Network Analysis, Complex Networks, Computational Social Science, Data Science, and Fairness. She is interested in designing machine learning and deep learning based methods for complex network data. Her current research is focussed on understanding inequalities in complex networks and algorithmic fairness in network and data science.

Authors and Affiliations

Simran Setia[1] · Anamika Chhabra[2] · Amit Arjun Verma[3] · Akrati Saxena[4]

Simran Setia
simran.setia@thapar.edu

Anamika Chhabra
anamika@study.iitm.ac.in

Amit Arjun Verma
mt4descentis@gmail.com

1 Department of Computer Science and Engineering, Thapar Institute of Engineering and Technology, Patiala, India

2 IIT Madras, Chennai, Tamil Nadu, India

3 CoreCLM, Seattle, Washington, USA

4 Leiden Institute of Advanced Computer Science, Leiden University, Niels Bohrweg 1, Leiden 2333CA, South Holland, The Netherlands