Computer Science Journal of Moldova, vol.**30**, no.1(88), 2022

Revisiting the Role of Classical Readability Formulae Parameters in Complex Word Identification (Part 2)*

Gayatri Venugopal, Dhanya Pramod †,
Jatinderkumar R. Saini ‡

Abstract

Accessibility of text is an attribute that deserves the attention of researchers and content creators. This study is an attempt to determine the lexical features that play a key role in identifying complex words in Hindi text. As the first step, we studied the parameters used in readability metrics in different languages and tested their importance on classifiers built on datasets created with the help of a user study. In part of the study, we reported the results of two different approaches used to label a word as complex. In this part, we compare the previous results with the results obtained from a third labeling approach. We found satisfactory evidence for certain parameters and also observed a new parameter that could be used while devising readability metrics for Hindi.

Keywords: complex word identification, readability, hindi, binary classification, natural language processing.

MSC 2010: 68R10, 68Q25, 05C35, 05C05.

©2022 by CSJM; G. Venugopal, D. Pramod, J.R. Saini

* This work was supported by Symbiosis Centre for Research and Innovation, Symbiosis International (Deemed University), grant number 1591

† Equal contribution

‡ Equal contribution

1 Introduction

Text simplification refers to the process of modifying a text in such a manner that it becomes more comprehensible to reader with no loss of information. The words that the content creators use may not always be understood by the consumers of the content. They may use words that they are comfortable with or that would distinguish them from the others [1]. Readable texts would not only help readers who are new to the language, but could also help readers with reading disabilities such as dyslexia [2], aphasia [3], and also readers with a poor level of literacy, and children [4]. Various readability formulae have been devised by researchers that would help identify the complexity level of a given text. We can find the usage of these formulae even today in various studies [5]–[7]. This study is focused on identifying the similarities between the lexical parameters used in the readability formulae, and the features deemed to be significant in determining a word in Hindi to be complex or not. Hindi, the official language of India, is ranked high in the list of languages spoken by first language speakers in the world [8]. Although the readability formulae have been devised for a multitude of languages, most of the formulae are centered around English. Besides readability formulae, we have also taken into account the lexical characteristics used in the creation of word lists that contain words that are considered to be simple.

We began our study by identifying the quantifiable characteristics of words that were used as parameters in commonly used readability formulae. We focused our study on the most common characteristics and used them as features of words in different classification models. The study focuses on the lexical complexity of a word. We used two approaches to identify the important features, the outcomes of which prominently highlighted the importance of the frequency of a word and the irrelevance of the word's length. This result contradicts the importance given to the length of a word by various readability formulae. We also observed that the number of hyponyms of a word, which has not been covered by any formula, is a key predictor of the complexity of a word.

This study is the second part of a two-part study that was conducted to determine the significance of parameters used in readability formulae, in the identification of complex words in Hindi text. In the earlier study, we used two approaches to create the dataset. The first approach consisted of deeming a word as complex if at least two users rated the word as complex, whereas a word was deemed complex using the second approach if the majority of the raters rated the word as complex. We discarded the first approach based on the results discussed in Part 1 of this study.

In this paper, we discuss another approach used to label a word as complex and report a comparison of the results obtained using this approach and the previous approach discussed in Part 1 of the study.

2 Lexical Parameters

This section gives an overview of the lexical parameters used in popular readability formulae and word-lists.

One of the earliest studies on readability in English was done by L.A. Sherman, an English professor, who claimed that the length of a sentence and the concreteness of a word impact the comprehension of the reader [9]. Although we could not find any word list or formula proposed by Sherman, there exist numerous studies that have proposed various parameters to test the readability of a given text, most of them written in English.

Rubakin was a prominent Russian writer who published a list of 1500 words in Russian in 1889, that he claimed to be easy to understand. According to Rubakin, words that were not known to the users and sentences that have many words act as hindrances to comprehension [10]. One of the popular word lists is the Teacher's Word Book, that was created in 1921 [11], wherein the author focussed on including words with a short length and high frequency. More words were added to this list in 1944, making it a list of 30,000 words [10].

Readability of text encompasses various aspects such as content, style, format and structure [12]. Our study is focused on the lexical parameters used in readability formulae. The Flesch Reading Ease

formula takes into account the number of syllables, number of words and number of sentences [13]. Researchers have adapted this formula to languages other than English by modifying the values of coefficients [14]–[16]. The number of syllables has been considered to be a major factor in various other readability measures such as the Gunning Fog index [17], the readability system created by Edward Fry [18], SMOG [19], and the Flesch-Kincaid Reading Ease formula [20]. Other common parameters used in readability formulae are word length [21], [22] and frequency [14], [23], [24].

In [25] the authors proposed a readability formula that focused on the number of words in a sentence and the number of sentences in 300 words in a given text. In [26] the authors devised a formula for determining the readability of Vietnamese text which took into account the average sentence length in characters, the average of word length in terms of the number of characters, and the percentage of difficult words that was calculated using a list of easy words.

The parameters used in readability formulae that targeted the Hindi language were length, number of consonants, and number of consonant conjuncts [27], [28]. Therefore our study aimed to analyse these lexical parameters and their importance in complex word classification models.

3 Methodology

We conducted a user study consisting of 50 native and 50 non-native speakers of Hindi in the age group of 18 to 30 years. The user study involved two steps – annotating complex words in a set of 100 sentences, and ranking the annotated word in comparison with its synonyms, using a Liker scale, where 1 indicated very complex and 5 indicated very simple.

3.1 Labeling Method

In order to build a classifier, our next step was to label a word as complex or simple. We used three approaches, two of which have been discussed in Part 1 of this study, which also contains the details of the

sources of feature values used to build the model. As part of our third approach, we labeled a word as complex if the average rating assigned to the word was less than or equal to 3.

In this approach, which we believe is more bias-free as compared to the first approach, only words that were rated by at least two participants were considered. The dataset thus generated consisted of 7326 records out of which 2958 records were labeled as complex and 4368 records were labeled as simple. The dataset size is less as compared to the dataset generated in approach 1 because we did not consider words that were ranked by only one person (in order to avoid bias). The training-test proportion was 70:30. The training set consisted of 5129 records out of which 40.38% were labeled 1 and 59.62% were labeled 0. Since this was not an undesired proportion, resampling techniques were not used.

3.2 Feature Evaluation Methods

As discussed in Part 1 of the study, we used the traditional as well as ensemble classification algorithms with k-fold cross-validation with five splits. The algorithms used were decision tree, support vector classifier, nearest centroid classifier, random forest, extra trees, Ada boost, gradient boosting and XG boost. We chose sense-normalised values of length, number of syllables, frequency of the lemma of the word, number of consonants, number of vowels, number of consonant conjuncts, number of synsets, number of synonyms, number of hypernyms and the number of hyponyms as the features used to build the classifiers.

In order to evaluate the features, we used two methods as shown in Figure 1 and Figure 2, respectively.

In the first method, eight models were built and each feature's importance value was calculated using permutation feature importance and exhaustive feature selection. The importance value generated from all the models are aggregated to obtain one score for each feature. Accuracy and Macro-F1 scores were used as the metrics to evaluate the performance.

In the second method, we tuned the models using random search hyperparameter tuning based on Receiver Operating Characteristic

(ROC) scores and built a soft voting classifier as our final model. The feature importance values were calculated for this model.

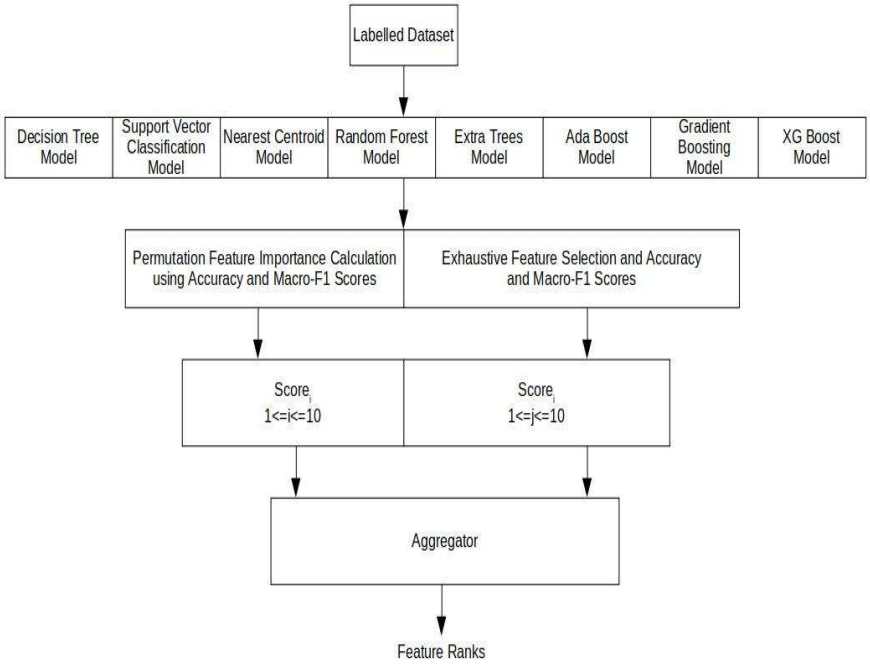

Figure 1. Method 1

4 Results and Discussion

Besides the features mentioned in the previous section, we included four other lexical features, which were synonyms, number of synsets, number of hyponyms and number of hypernyms to build the models. We calculated the feature importance scores using accuracy and macro-F1 scores, and assigned the highest rank, e.g. 1, to the feature with the highest average. With regard to exhaustive feature selection, the value 1 was assigned to a feature if it was present in a feature subset for a model, and the value 0 was assigned to it if it was not present in

the feature subset. The results can be seen in Table 1. The ROC curve for the models can be seen in Figure 3.

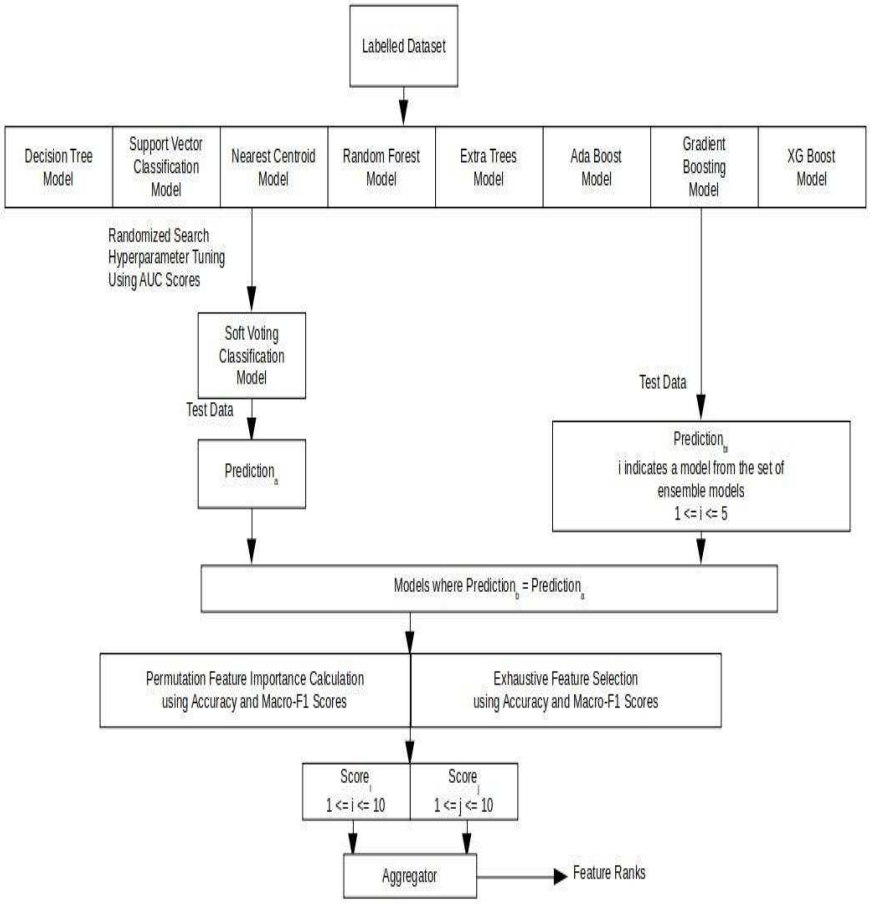

Figure 2. Method 2

Table 1. Feature importance values for each feature based on accuracy and macro-F1 scores for all the models

Feature	**Permutation Feature Importance**		**Exhaustive Feature Selection**	
	Accuracy	**Macro-F1**	**Accuracy**	**Macro-F1**
n_synonyms	5.125	5.75	0	0
n_synsets	4.625	4.625	0	0
frequency	10	10	0.625	0.625
n_hyponyms	8.25	8	0	0
n_syllables	3	2.75	0	0
n_hypernyms	4.75	5	0	0
length	4.125	4.125	0	0
n_consonants	6.5	6.5	0	0
n_vowels	3.75	3.5	0	0
n_consonantconjuncts	4.875	4.75	0	0
Mean	5.5	5.5	0.0625	0.0625
Median	4.8125	4.875	0	0
Standard Deviation	2.157	2.1755	0.1976	0.1976

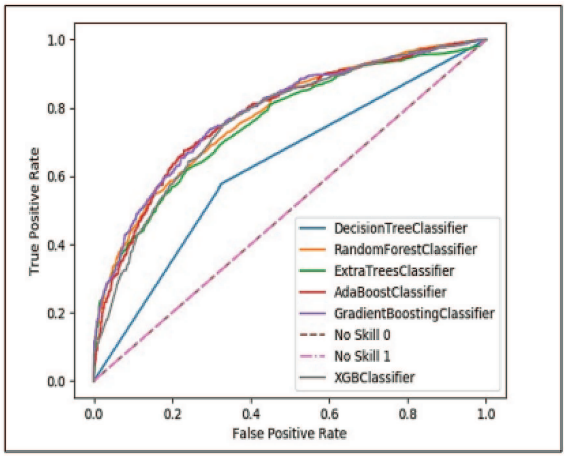

Figure 3. ROC curve for all the tree based models

We then took an aggregate of the values for each method of feature importance calculation. The values can be seen in Table 2.

Table 2. The aggregate of the feature importance values for each feature for all the models

Feature	**Importance Value Based On**		**Feature Importance Value**
	Accuracy	**Macro-F1**	
n_synonyms	5.125	5.75	5.4375
n_synsets	4.625	4.625	4.625
frequency	10.625	10.625	10.625
n_hyponyms	8.25	8	8.125
n_syllables	3	2.75	2.875
n_hypernyms	4.75	5	4.875
length	4.125	4.125	4.125
n_consonants	6.5	6.5	6.5
n_vowels	3.75	3.5	3.625
n_consonantconjuncts	4.875	4.75	4.8125

The second method encompasses analysing the importance values obtained for all the tree-based models. These values were compared against the baselines (ALL 0 and ALL 1). The results can be seen in Table 3.

As was the case with Approach 2 mentioned in Part 1, ensemble classifiers showcase a better performance than the traditional decision tree classifier. We then built a soft voting classifier and selected the label with the maximum vote, among the labels generated using the ensemble classifiers and the tuned ensemble classifiers. The Area under the ROC Curve (AUC) scores can be seen in Table 4.

In order to identify the significant features, we considered each prediction and used the classification models which made the correct prediction. The feature importance values of these models were calculated by adopting the strategy implemented in Method 1. The ranks of the features calculated based on their importance values, can be seen in Table 5.

Table 3. Analysis of Importance Values

Classifier	**Macro-F1**	**Accuracy**
Baseline (ALL 0)	0.187	0.596
Baseline (ALL 1)	0.144	0.404
Support Vector	0.289	0.618
Nearest Centroid	0.307	0.611
Extra Trees	0.338	0.695
Random Forest	0.348	0.714
XGB	0.343	0.703
Gradient Boosting	0.353	0.719
Ada Boost	0.353	0.719
Decision Tree	0.319	0.652

Table 4. AUC Scores

Model	**AUC Score**
Ada	0.776
Tuned Ada	0.781
Extra Trees	0.760
Tuned Extra Trees	0.762
Gradient Boosting	0.783
Tuned Gradient Boosting	0.755
Random Forest	0.770
Tuned Random Forest	0.785
XGBoost	0.785
Tuned XGBoost	0.782
Soft Voting	0.790

The ranks obtained using this approach are identical to the ranks obtained using Approach 2 in Part 1 of the study. Therefore we have strong evidence that frequency is a major predictor of the complexity of a word. This aligns with the results reported by studies conducted on other non-Indian languages.

Table 5. Feature Importance Values and Ranks

Feature	**Feature Importance Value**	**Rank**
frequency	1.156	1
n_hyponyms	1.04	2
n_syllables	1.036	3
n_vowels	1.033	4
n_consonants	1.027	5
n_synsets	1.023	6
n_hypernyms	1.022	7
length	1.017	8
n_synonyms	1.013	9
n_consonantconjuncts	1.002	10

5 Conclusion and Future Scope

The goal of the study was to ascertain the significance of lexical parameters used in readability measures, in complex word identification in Hindi text. We used two methods and three approaches to approach the problem. The feature importance values were calculated using accuracy and macro-F1 scores. A soft voting classifier was used as it performed better than the individual models involved in the study.

Through this study, we reinstated the importance of the role that frequency plays in determining the complexity of a word. Many readability measures used word length as one of the parameters. However, we found from both our approaches, Approach 2 and 3, that length of a word is not a significant factor. The readability measures also focused on the number of syllables, which was proven to be an important predictor of word complexity. We suggest the use of the number of hyponyms of a word as a parameter in a readability measure for Hindi text as this has proven to be an important feature in a complex word classifier.

This work could be refined by using a different approach for complex

word labeling and by tuning the models using grid search hyperparameter tuning. Researchers may use this word to create a readability metric for Hindi text focusing on lexical attributes of the text.

6 Acknowledgements

This study was sponsored by Symbiosis Centre for Research and Innovation, Symbiosis International (Deemed University) and has approval from the Independent Ethics Committee, Symbiosis International (Deemed University) in July 2019. We are grateful to the Linguistic Data Consortium for Indian Languages and IIT Bombay for providing the corpora that were used in the study, and also the Stanford NLP group for releasing the stanfordnlp Python package that was used to retrieve the lemmas of words in Hindi [29]. We thank the reviewers for their valuable comments that helped improve the paper.

References

[1] E. L. Thorndike, "The psychology of semantics," *The American journal of psychology*, vol. 59, no. 4, pp. 613–632, 1946.

[2] L. Rello, R. Baeza-Yates, S. Bott, and H. Saggion, "Simplify or help?" in *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility - W4A '13*. ACM Press, 2013. [Online]. Available: https://doi.org/10.1145/2461121.2461126

[3] J. Carroll, G. Minnen, Y. Canning, S. Devlin, and J. Tait, "Practical simplification of english newspaper text to assist aphasic readers," in *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, 1998, pp. 7–10.

[4] J. De Belder and M.-F. Moens, "Text simplification for children," in *Prroceedings of the SIGIR workshop on accessible search systems*. ACM; New York, 2010, pp. 19–26.

[5] M. A. Kugar, A. C. Cohen, W. Wooden, S. S. Tholpady, and M. W. Chu, "The readability of psychosocial wellness patient resources: improving surgical outcomes," *Journal of Surgical Research*, vol. 218, pp. 43–48, 2017.

[6] J. C. Brewer, "Measuring text readability using reading level," in *Encyclopedia of Information Science and Technology, Fourth Edition*. IGI Global, 2018, pp. 1499–1507. [Online]. Available: https://doi.org/10.4018/978-1-5225-2255-3.ch129

[7] J. M. Marsh, T. D. Dobbs, and H. A. Hutchings, "The readability of online health resources for phenylketonuria," *Journal of Community Genetics*, vol. 11, no. 4, pp. 451–459, mar 2020. [Online]. Available: https://doi.org/10.1007/s12687-020-00461-9

[8] "Languages of the world." [Online]. Available: www.ethnologue.com

[9] L. A. Sherman, *Analytics of literature: A manual for the objective study of English prose and poetry*. Ginn, 1893.

[10] E. L. Thorndike and I. Lorge, "The teacher's word book of 30,000 words." 1944.

[11] E. L. Thorndike, "The teacher's word book," 1921.

[12] W. S. Gray and B. E. Leary, "What makes a book readable." 1935.

[13] R. Flesch, "A new readability yardstick." *Journal of Applied Psychology*, vol. 32, no. 3, pp. 221–233, 1948. [Online]. Available: https://doi.org/10.1037/h0057532

[14] S. Štajner and H. Saggion, "Readability indices for automatic evaluation of text simplification systems: A feasibility study for spanish," in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 2013, pp. 374–382.

[15] W. Douma *et al.*, "Readability of dutch farm papers: a discussion and application of readability-formulas." *Readability of Dutch farm papers: a discussion and application of readability-formulas.*, no. 17, 1960.

[16] R. Brouwer, "Onderzoek naar de leesmoeilijkheden van nederlands proza," *Pedagogische studiën*, vol. 40, pp. 454–464, 1963.

[17] R. Gunning *et al.*, "Technique of clear writing," 1952.

[18] E. Fry, "A readability formula that saves time," *Journal of reading*, vol. 11, no. 7, pp. 513–578, 1968.

[19] G. H. Mc Laughlin, "Smog grading-a new readability formula,"

Journal of reading, vol. 12, no. 8, pp. 639–646, 1969.

[20] J. P. Kincaid, J. Fishburne, R. R. P., C. R. L., and B. S., "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel," Tech. Rep., feb 1975. [Online]. Available: https://doi.org/10.21236/ada006655

[21] R. Senter and E. A. Smith, "Automated readability index," CINCINNATI UNIV OH, Tech. Rep., 1967.

[22] M. Coleman and T. L. Liau, "A computer readability formula designed for machine scoring." *Journal of Applied Psychology*, vol. 60, no. 2, pp. 283–284, 1975. [Online]. Available: https://doi.org/10.1037/h0076540

[23] G. R. Weir and C. Ritchie, "Estimating readability with the strathclyde readability measure," in *ICT in the Analysis, Teaching and Learning of Languages, Preprints of the ICTATLL Workshop 2006*, 2006, pp. 25–32.

[24] A. Anula, "Tipos de textos, complejidad lingüıstica y facilicitación lectora," in *Actas del Sexto Congreso de Hispanistas de Asia*, 2007, pp. 45–61.

[25] N. Hazawawi, M. Zakaria, and S. Hisham, "Formulating an algorithm to detect readability level of malay texts," *Proceedings of Mechanical Engineering Research Day 2017*, vol. 2017, pp. 77–78, 2017.

[26] A.-V. Luong, D. Nguyen, and D. Dinh, "A new formula for vietnamese text readability assessment," in *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*. IEEE, nov 2018. [Online]. Available: https://doi.org/10.1109/kse.2018.8573379

[27] M. Sinha, S. Sharma, T. Dasgupta, and A. Basu, "New readability measures for bangla and hindi texts," in *Proceedings of COLING 2012: Posters*, 2012, pp. 1141–1150.

[28] M. Sinha, T. Dasgupta, and A. Basu, "Text readability in hindi: A comparative study of feature performances using support vectors,"

in *Proceedings of the 11th International Conference on Natural Language Processing*, 2014, pp. 223–231.

[29] P. Qi, T. Dozat, Y. Zhang, and C. D. Manning, "Universal dependency parsing from scratch," *arXiv preprint arXiv:1901.10457*, 2019.

Gayatri Venugopal, Dhanya Pramod,
Jatinderkumar Saini

Received May 27, 2021
Accepted September 9, 2021

Gayatri Venugopal
Symbiosis Institute of Computer Studies and Research (SICSR)
Symbiosis International (Deemed University) (SIU),
Model Colony, Pune, Maharashtra, India
Phone: +91-9665856569
E-mail: gayatri.venugopal@sicsr.ac.in

Dhanya Pramod
Symbiosis Centre for Information Technology (SCIT)
Symbiosis International (Deemed University) (SIU),
Hinjewadi, Pune, Maharashtra, India
E–mail: dhanya@scit.edu

Jatinderkumar R. Saini
Symbiosis Institute of Computer Studies and Research (SICSR)
Symbiosis International (Deemed University) (SIU),
Model Colony, Pune, Maharashtra, India
E-mail: saini_expert@yahoo.com