Computer Science Journal of Moldova, vol.**29**, no.3(87), 2021

Analyzing Complex Words in Hindi using Parameters of Classical Readability Formulae (Part 1)*

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

Abstract

Readability of a passage indicates the extent to which the meaning of the text can be understood; this could be represented in terms of the age that person should be of, or the grade that a person should be in, to understand the text. Numerous word lists and readability formulae have been devised by researchers who tested the readability of texts by involving children and adults. Most of these resources have been built for the English language. This study aims to analyse the complex words in Hindi sentences that were derived from a Human Intelligence Task (HIT), using variables considered in the widely adopted readability measures that focus on the lexical aspects of a sentence. Although there have been studies that analyse the readability of texts, this study claims to be the first of its kind, that aims to determine whether the parameters of traditional readability measures contribute significantly to context-agnostic models that classify a Hindi word as complex or simple. We report the results of two approaches used to deem a word as complex and determine the best approach out of the two. The model built using this approach was used to identify the most significant features.

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

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

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

† Equal contribution

‡ Equal contribution

1 Introduction

Comprehension is an essential characteristic of text that often goes unnoticed by content creators. In order to help content creators develop content according to the use base of the text, researchers have devised various readability formulae, that indicate the age or the grade of the target reader. Though these measures are widely available for English, and have been adapted for use in other languages, the use of these formulae to analyse the complexity of words in Hindi text is yet to be explored. Since all such measures work at a text level, we cannot use them directly to assess the complexity of a word. Similarly, the word lists developed by researchers in the area focus on English and a select group of target users, such as children belonging to a particular grade. No such word lists are available for Hindi. Therefore we aim to study the basis of selection of words in the widely popular word lists in English and a few non-English word lists in order to find similarities among the selection criteria.

Although we focus on readability formulae in this paper, we do not disregard the role that familiarity plays in the comprehensibility of text. [1] explains how the text itself can be considered the centre of attention, i.e., all the components in the text should be analysed including words, phrases and sentences. The researcher also showcases how the comprehensibility of the text is closely associated with the cognitive ability as well as vocabulary and experience of the user. Along the same lines, [2] reported the influence of properties of text and the reader's understanding of the topic. [3] studies the comprehensibility of text using the Karlsruhe comprehensibility concept, which encompasses elements such as perceptibility, simplicity, correctness, structure, motivation, and concision. Hence, readability and comprehensibility are equally significant to understand the meaning of a text. However, the scope of our study is limited to the parameters of readability measures. Although significant amount of work has been done in the field of complex word identification in languages other than Hindi, we choose not to consider the parameters proposed by those studies. Our research objective is to evaluate the parameters used in word lists and formulae that have been prevailing since the early 1920s, on Hindi words

identified as complex by users in a Human Intelligence Task (HIT).

This paper consists of the following sections: Section 2 highlights the quantifiable parameters of word lists and readability formulae prevalent since 1921 across languages. Section 3 consists of the methodology used to annotate complex words, a description of the dataset and the approaches used to identify the key features that play a role in determining whether a word is complex or not given a Hindi sentence. Section 4 contains the results and interpretation. Section 5 contains the conclusion, limitations and future scope.

2 Readability Measures and Existing Work

This section has been divided into two sub-sections, dedicated to word lists and readability formulae, respectively.

2.1 Word Lists

One of the word lists that was widely used, was published in the Teacher's Word Book in 1921 [4], wherein the author claimed that a shorter word is easier to understand as compared to a longer word. Thorndike considered frequency to be a major factor to judge complexity. The list consists of 10,000 English words, that are grouped by frequency and associated with a grade level between kindergarten and 12. This list was used by researchers working in the English language to create a readability formula [5]. The formula was dependent on the word list and the number of unique words in the text. It was observed that this list consists of various forms of one word, i.e., the lemmas of the words were not considered while creating the list. The list also consists of many proper nouns. Since proper nouns have no alternative words, their inclusion in the list may not be utilised while simplifying words. In 1941, Edgar Dale created a wordlist using Thorndike's list and the International Kindergarten Union list, based on the frequency of the words [6]. This list was targeted at children in the first grade.

2.2 Readability Formulae

Researchers have developed readability scores which indicate the level of education that the reader would need in order to read the content

easily. The scope of our study however, is restricted to the use of word-level parameters in readability formulae.

Flesch proposed a reading ease formula in 1948 [7]. He also used the McCall-Crabbs Standard Test for English to assess comprehension [8]. Three years later, in 1951, [9] modified this formula and created a simplified Flesch reading ease formula. In 1952, Gunning devised the Gunning Fog index, that was based on the McCall-Crabbs lessons [10], [11]. The Automated Readability Index was created and tested in 1967 [12], in which the researchers found a strong correlation between the index and the grade level of the text. The index represents the approximate age that a reader should be of, to understand the text. In 1968, Edward Fry devised a graph-based system to calculate the grade level of a given text [13]. The graph targets the English language in the context of the US education system. The average number of syllables and the average number of sentences for every hundred words are plotted on a graph. The grade is then calculated from the graph. In the subsequent year, McLaughlin devised a new readability formula and named it SMOG [14]. He claimed that the complexity of a text depends on the number of polysyllabic words in the text. Words having three or more syllables were considered for inclusion in the formula. The reading grade produced from the formula is based on the reading grades calculated from the Thorndike-McCall Reading Test. This formula produces as result, the number of years of education the reader is required to have, to understand the text. 1 indicates that the readability level is low, and 12 indicates that the text has a high readability level. Coke and Rothkopf, in 1970, worked on calculating the syllables of a word with the help of mathematical evidences [15]. In 1975, the Flesch formula was modified for English text and called the Flesch-Kincaid Reading Ease formula [16]. The result of the formula represents the American school grade that the reader should be in, in order to understand the text. In the same year, another formula was devised by [17], the output of which represents the grade level of the text. In 2006, Weir and Ritchie proposed the Strathclyde readability measure that focused on the frequency of a word [18]. Anula in 2007, suggested a Lexical Complexity Index, that used the lexical density and index of words with low frequency as its parameters [19], [20].

2.3 Readability Studies in Hindi

The earliest study on readability in Hindi was initiated by [21], who tested the readability of short stories in Hindi using English readability formulae. [22] used the traditional readability formula devised for English, on Hindi text and found that the readability of a given text cannot be predicted by surface features alone. [23] proposed a model to assess the readability of text in Hindi using the average number of consonants and the number of consonant conjuncts in the text. Another observation made by [24] in their study, was that the number of consonant conjuncts and average word length are good predictors of text readability.

We could not use the proposed formulae as we are not considering the readability at the sentence or the text level. However, we aim to find a correlation between the word-level parameters that are common across these formulae, and the complexity of the word, that was obtained by conducting a human intelligence task. A summary of the parameters that were considered while creating the wordlists and readability formulae can be seen in Tables 1 and 2, respectively.

From Table 1 and Table 2, we chose the parameters that are quantifiable – length, frequency, number of syllables, number of vowels, number of consonants and number of consonant conjuncts. The research objective of the study was to determine the relationship between these parameters and word complexity in Hindi, where complexity refers to the understandability of a word by a reader.

Table 1. Word level parameters used in word lists

Language	**Word Level Parameters**
Russian	Familiarity of the word Rubakin in 1889 [25]
English	Length, Frequency [4]
English	Length, Frequency [6]

Table 2. Word level parameters used in readability formulae

Language	**Word Level Parameters**
English	Length, Frequency [5]
English, Dutch	Number of syllables [7]
English	Number of syllables [9]
English	Number of syllables [10]
English	Length [12]
English	Number of syllables [13]
English	Number of syllables [14]
English	Number of syllables, number of vowels [15]
English	Number of syllables [16]
English	Length [17]
English	Frequency [18]
English	Frequency [19]
Hindi	Length, number of consonants, number of consonant conjuncts [23], [24]
Vietnamese	Length [26]

3 Methodology

Our objective was to determine whether p_i contributes significantly to context-agnostic binary classification models that categorise a word as simple or complex, where p indicates the set of parameters {length, frequency, number of syllables, number of vowels, number of consonants, number of consonant conjuncts} and i indicates the position of a parameter in the set with its value ranging from 1 to 6. Henceforth these parameters would be referred to as features in the context of machine learning.

In order to create the dataset, we conducted a Human Intelligence Task (HIT) study. 100 native and non-native speakers of Hindi in the age group between 18 and 30 years participated in the study. Each participant was required to complete two tasks. In task 1, the participant was presented with a sentence on the screen. They highlighted the

word/s whose meaning/s they did not understand. They were asked to highlight the word even if they guessed the meaning from the context, without knowing the meaning of the word. In the subsequent task, task 2, they were presented with a set of words (synset of the word they marked as complex in task 1). They assigned a rank to every word – 1 being complex and 5 being simple. The values of features of the words were retrieved using the Hindi WordNet [27]. The frequency was calculated from a corpus that we created by collating texts from novels and stories as well as from other existing corpora [28], [29].

Initially we determined the correlation of the individual features of the words with the complexity. But correlation could be misleading as a particular feature may not have any significant correlation, but a set of features together may be correlated with the target. We used permutation feature importance and exhaustive feature selection to identify the important features. Permutation feature importance shuffles the value of a feature and determines how much this change affects the error in the model. Exhaustive feature selection is used to select a subset of features that give the best performance. Since 5-fold cross-validation was used, the average values across the 5 folds were considered for each feature for the models built using classical machine learning models, the details of which can be seen below. We chose a context agnostic methodology as the readability formulae do not focus on the context of a word. Since the complexity of a word is subjective in nature, we used and tested multiple methods in order to deem a word to be complex. We used three approaches to assign the label 1 (complex) to a word. The approaches and a description of the datasets generated using the approaches are as follows.

3.1 Approach 1

In the first approach, a word was deemed as complex if at-least two raters annotated the word as complex in task 2. We chose this number as 2 in order to avoid any bias. The dataset consists of 11565 records with 2183 complex labels and 9382 simple labels. The training and test sets were created using a 70-30 proportion. The training set distribution contained 8095 records out of which 18.88% were labeled 1

(complex) and 81.12% were labeled 0 (simple). Owing to the imbalance, the training was done on a resampled dataset. Both, oversampled data and undersampled data were used for feature extraction. Oversampling was performed using SMOTE, and NearMiss-3 was used for undersampling.

3.2 Approach 2

We considered using percentage of observed agreement to determine the agreement for every word. We could not use the standard inter-annotator agreement methods as they did not assess the agreement for one item. We selected only those words for which the observed agreement was >= 75%. This led to a considerable amount of data loss as the agreement was very low owing to varying backgrounds of the participants. Therefore we used the majority vote as the label. The size of the dataset created using Approach 2 was 10499. After deleting the records in which there was a tie, we were left with 8576 records. We removed the words whose information was not present in the Hindi WordNet. The records in which the word was labeled by only one annotator were not considered. The final dataset consisted of 6154 records with 3260 complex word records and 2894 simple word records. Similar to the other approaches, the train-test ratio was 70:30. The training set consisted of 4308 records, out of which 53% were labeled complex and 47% were labeled simple. The dataset was balanced, therefore resampling was not required.

3.3 Classifiers and Evaluation Metrics

We considered eight models for classification, out of which five were ensemble models. The models used to classify the data were decision tree, support vector classifier, nearest centroid classifier, random forest, extra trees, Ada boost, gradient boosting and XG boost. k-fold cross-validation with five splits was used for classification. Collinearity was observed among the features, hence we chose not to use logistic regression. The features used to generate the models were the following: 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.

The values of all the features were normalized between 0 and 1 by using minmax normalisation. The values for only the words in the synset were considered while normalising, as it did not seem intuitive to compare the values of features of unrelated words. We did not compare the features of unrelated words. We created groups of words, each group indicating a synset. After obtaining values of features of words belonging to the same synset, we normalized those values using MinMax normalization. By normalizing all the values in all the synsets, based on the range of values in each synset, to a value in the range of 0-1, we made it possible to compare the values of features of different synsets. The common limitation of min-max normalization is that it is sensitive to outliers. But since we require our values to be normalized depending on the values of the outliers as well, we do not consider this as a limitation. We then created our final dataset that consisted of the lexical features, frequency, and the label of a word. Synsets in which a feature had identical values, i.e., there was no minimum or maximum, were excluded as they would not contribute toward building the model. In order to evaluate the features, we used two methods.

In the first method, the models were trained with default parameters. The feature importance values were calculated using permutation feature importance and exhaustive feature selection. Permutation feature importance was calculated for each model as opposed to feature importance. It refers to the change in the score of the model when each feature is randomly shuffled, one at a time. Exhaustive Feature Selection was used to find the relevant feature set for each model. The evaluation metrics used were accuracy and macro-F1. We calculated the permutation feature importance value for a feature by averaging the permutation importance values across all the folds in the cross validated model. The intersection of the values was obtained from the features generated using exhaustive feature selection, from the oversampled and undersampled data. The permutation importance values for features obtained from oversampled data and undersampled data were combined and used as the feature importance values.

In the second method, we used soft voting classification and random search hyperparameter tuning of the models to identify the most significant features. Receiver Operating Characteristic (RUC) scores were used to tune the models. Precision-Recall curves were used in Approach 1 owing to imbalanced datasets. The values were compared against the baselines – ALL 0 and ALL 1. Permutation feature importance and exhaustive feature selection were used, as in Method 1, based on accuracy and macro-F1 scores for calculating the feature importance values.

To summarise this section, we provide the following descriptions of the strategies and the terms considered as part of the study:

- Labeling Approach
  Approach 1 – A word is labeled complex if at-least two annotators rated it 3 or less
  Approach 2 – A word is labeled complex if the majority rating it received is 3 or less
- Models
  Traditional Classifiers – Decision Tree, Nearest Centroid, Support Vector Classifier
  Ensemble Classifiers – Random Forest, Extra Trees, Ada Boost, Gradient Boosting, XG Boost
  Soft Voting Classifier
- Methods
  Method 1 – Feature importance values are extracted from models trained with default hyperparameters
  Method 2 – Feature importance values are extracted from tuned ensemble models along with a voting classifier
- Features
  The features that were present in the training data were number of synonyms, number of synsets, frequency, number of hyponyms, number of syllables, number of hypernyms, length, number of consonants, number of vowels, number of consonant conjuncts.

4 Results and Discussion

After implementing Method 1, the results were recorded for each of the eight models with resampling for Approach 1, and without resampling for Approach 2. Although we selected six features in this paper, we report the results for more features that we included in our study, based on the findings from the literature. These features were number of synonyms, number of synsets, number of hyponyms and number of hypernyms.

At the end of the experiment, we achieved the following results:

- Permutation feature importance values calculated using accuracy score and macro-F1 score for each model under each approach

  The feature with the highest average was assigned the highest rank, where rank indicates importance of the feature. The average of the ranks for each feature across all the models was calculated separately for the ranks obtained by using the accuracy score and for the ranks obtained by using the macro-F1 score.
- A feature subset for each model obtained using exhaustive feature subset selection under each approach calculated using the accuracy score and the macro-F1 score.

A feature was assigned the value 1 if it was present in a feature subset for a model and 0, otherwise. The average of the values for each feature across all the models was calculated separately for the values obtained by using the accuracy score and for the values obtained by using the macro-F1 score.

The results can be seen in Tables 3 and 4 for Approaches 1 and 2, respectively.

For each approach, we combined the values generated using permutation feature importance and exhaustive feature subset methods based on accuracy and macro-F1 separately. The mean of the accuracy values and macro-F1 values was taken for each feature in each approach. The values can be seen in Tables 5 and 6 for Approaches 1 and 2, respectively.

Table 3. Feature importance values for each feature based on accuracy and macro-F1 scores for all the models in Approach 1

Feature	**Permutation Feature Importance**		**Exhaustive Feature Selection**	
	Accuracy	**Macro-F1**	**Accuracy**	**Macro-F1**
n_synonyms	8.375	8.375	0.375	0.375
n_synsets	7.5	7.625	0.625	0.5
frequency	7.375	7.5	0.25	0.25
n_hyponyms	7	7	0.375	0.375
n_syllables	6.375	6.375	0.25	0.25
n_hypernyms	6	5.75	0.375	0.375
length	3.875	3.75	0.125	0.125
n_consonants	3.375	3.25	0.375	0.375
Mean	5.5	5.5	0.3125	0.3
Median	6.1875	6.0625	0.3125	0.3125
Standard Deviation	2.2024	2.2111	0.1473	0.1208

Table 4. Feature importance values for each feature based on accuracy and macro-F1 scores for all the models in Approach 2

Feature	**Permutation Feature Importance**		**Exhaustive Feature Selection**	
	Accuracy	**Macro-F1**	**Accuracy**	**Macro-F1**
n_synonyms	5.125	5.75	0	0
n_synsets	4.75	4.75	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	4.75	4.625	0	0
Mean	5.5	5.5	0.0625	0.0625
Median	4.75	4.875	0	0
Standard Deviation	2.1562	2.1755	0.1976	0.1976

Table 5. The aggregate of the feature importance values for each feature for all the models for Approach 1

Feature	**Importance Value Based On**		**Feature Importance Value**
	Accuracy	**Macro-F1**	
n_synonyms	8.75	8.75	8.75
n_synsets	8.125	8.125	8.125
frequency	7.625	7.75	7.6875
n_hyponyms	7.375	7.375	7.375
n_syllables	6.625	6.625	6.625
n_hypernyms	6.375	6.125	6.25
length	4	3.875	3.9375
n_consonants	3.75	3.625	3.6875
n_vowels	3.125	3.25	3.1875
n_consonantconjuncts	2.375	2.5	2.4375

Table 6. The aggregate of the feature importance values for each feature for all the models for Approach 2

Feature	**Importance Value Based On**		**Feature Importance Value**
	Accuracy	**Macro-F1**	
n_synonyms	5.125	5.75	5.4375
n_synsets	8.125	8.125	8.125
frequency	10.625	10.625	10.625
n_hyponyms	8.25	8	8.125
n_syllables	6.625	6.625	6.625
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	4.75	4.625	4.6875

We made the following observations during the process of calculating the importance values:

- Number of vowels has 0 effect on the accuracy of the model trained using XGB classifier in Approach 1.
- The length and the number of consonant conjuncts have 0 effect

on the macro-F1 value of the model trained using XGB classifier in Approach 1.

Owing to these observations that indicate the drastic changes in the importance values of one feature in different models, we devised Method 2. In Method 2, we analysed the accuracy, macro-F1 scores and the RUC scores for all the tree-based models. The values were compared against the baselines (ALL 0 and ALL 1). We will discuss the results of both the approaches in this subsection. The metrics for Approach 1 is given in Table 7.

Table 7. Metrics for Approach 1

Classifier	**Macro-F1**	**Accuracy**
Baseline (ALL 0)	0.224	0.811
Baseline (ALL 1)	0.079	0.189
Support Vector	0.255	0.466
Nearest Centroid	0.255	0.466
Extra Trees	0.282	0.652
Random Forest	0.282	0.652
XGB	0.266	0.588
Gradient Boosting	0.266	0.588
Ada Boost	0.266	0.588
Decision Tree	0.278	0.655

Table 7 contains the average of the evaluation metrics calculated from oversampled and undersampled data. There is no visible improvement using any model. As can be seen, the accuracy and macro-F1 scores are either lower than the baseline scores, or do not vary significantly from the baseline. The dataset created using Approach 1 was imbalanced, therefore we plotted a precision-recall curve, as is shown in Figure 1. We also compared it with the precision-recall curves for Approach 2, as is shown in Figure 2.

The plot for Approach 1 clearly indicates that there is no common point between precision and recall where any model achieves good results. Therefore, based on the results of all the models, we decided to discard Approach 1.

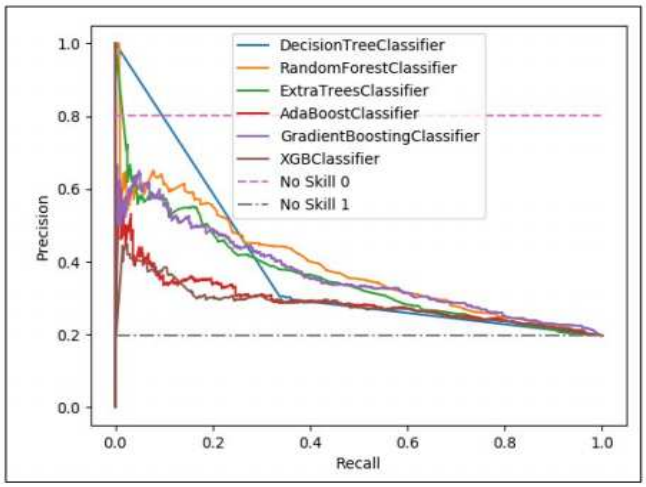

Figure 1. Precision-Recall curve for Approach 1 for all the tree based models

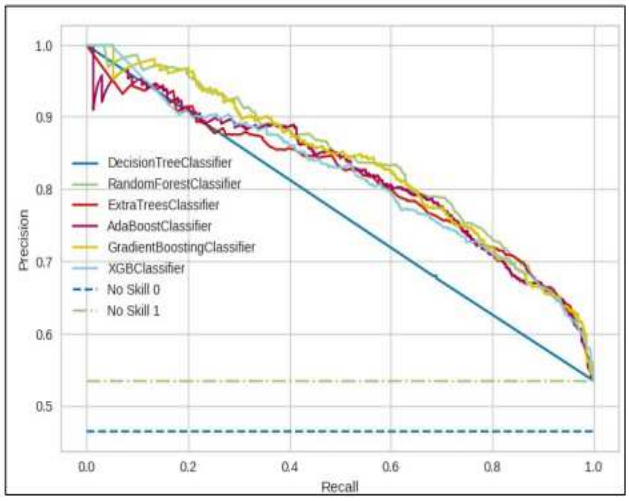

Figure 2. Precision-Recall curve for Approach 2 for all the tree based models

We observed that all the ensemble classifiers performed better than

the decision tree, although the difference between these classifiers is not huge.

The metrics for Approach 2 are given in Table 8, and the ROC curve can be seen in Figure 3.

Table 8. Metrics for Approach 2

Classifier	**Macro-F1**	**Accuracy**
Baseline (ALL 0)	0.16	0.47
Baseline (ALL 1)	0.173	0.53
Support Vector	0.289	0.618
Nearest Centroid	0.315	0.632
Extra Trees	0.359	0.718
Random Forest	0.363	0.727
XGB	0.366	0.733
Gradient Boosting	0.373	0.747
Ada Boost	0.369	0.738
Decision Tree	0.331	0.663

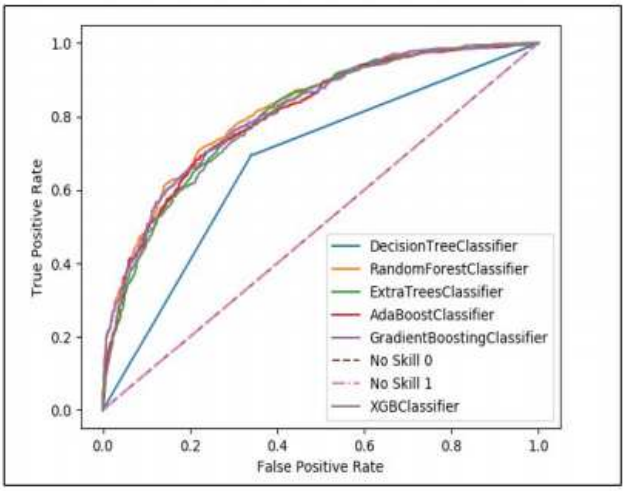

Figure 3. ROC curve for Approach 2 for all the tree based models

All the ensemble classifiers perform better than the decision tree,

although the difference between the plots of the ensemble classifiers is minimal.

We then used a soft voting process on Approach 2, where the probabilities of each class returned by all the models were averaged and the class with the highest average probability was kept. The label with the maximum vote, among the labels generated using the ensemble classifiers and the tuned ensemble classifiers was selected. We tuned the models using randomised search but observed that there is no significant difference between the results of the base model and the model obtained after performing the randomised search tuning. The Area under the ROC Curve (AUC) scores can be seen in Table 9.

Table 9. AUC Scores

Model	**Approach 2**
Ada	0.803
Tuned Ada	0.800
Extra Trees	0.801
Tuned Extra Trees	0.807
Gradient Boosting	0.812
Tuned Gradient Boosting	0.808
Random Forest	0.814
Tuned Random Forest	0.817
XGBoost	0.815
Tuned XGBoost	0.797
Soft Voting	0.824

As can be seen, the AUC score for the soft voting classifier was better than the individual scores of the classifier.

The next step was to identify the relevant features.

For every prediction, we used the classifiers which gave the correct prediction, and calculated the importance values of the features in the models. The importance values were derived using a combination of the exhaustive feature set and the ranks obtained from the permutation feature importance method. In order to calculate the importance value, the same strategy used in Method 1 was followed. The features were

arranged in decreasing order of their importance value, as can be seen in Table 10. The ranks have been displayed in Table 10.

Table 10. Features in decreasing order of importance value

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

We eliminated Approach 1 owing the inability of any model to produce a result better than the baseline.

As can be seen, frequency tops the list of features as the most important feature. Length, synonyms and consonant conjuncts are the features with the lowest importance values. The number of syllables and the number of vowels have approximately the same importance. The number of consonants, the number of synsets and the number of hypernyms fall among the features with lower importance values.

5 Conclusion and Future Scope

The objective of the study was to determine whether the parameters of classical readability formulae used to assess the readability of English as well as non-English text can be used to determine whether a word in a given Hindi sentence is complex or not. We used various methods to extract the important features from eight models, out of which five were tree-based ensemble models. A voting classifier was used as it produced better AUC scores as compared to all the individual models. We calculated the feature importance using accuracy and macro-F1 scores

within the permutation feature importance method and the exhaustive feature selection method. From the experiments, we observed that frequency is a common factor in readability of text as well as complexity of a word. However, length, which has been focused upon by most of the readability formulae does not rank high in the significant factors for complex word identification. The number of consonant conjuncts, which is a unique feature of Hindi words does not play a significant role in determining its complexity. However the number of syllables is an important feature. Another parameter that was not present in the readability formulae and word lists was the number of hyponyms. We observed that the number of hyponyms of a word, i.e., the number of words that are considered to be a sub-category of the target word, plays a key role in ascertaining the complexity of the target word.

In order to extend this work, we would recommend the use of grid search to tune the models more efficiently as compared to randomized search. We also suggest the addition of more lexical features to build models. Although we could not use grid search, we were able to achieve satisfactory results from the randomized search technique. It is evident from the results that there is a need for devising a word-level as well as a sentence level readability metric for Hindi.

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. We thank the reviewers for their valuable comments that helped improve the paper.

References

[1] S. Wolfer, "Comprehension and comprehensibility," *Translation and comprehensibility: Arbeiten zur Theorie und Praxis des*

Übersetzens und Dolmetschens, vol. 72, pp. 33–52, 2015.

[2] S. Hansen-Schirra, K. Maksymski, S. Wolfer, and L. Konieczny, "Investigating comprehensibility of german popular science writing," *Translation and comprehensibility*, vol. 72, p. 227, 2015.

[3] S. Göpferich, "Comprehensibility assessment using the karlsruhe comprehensibility concept," *The Journal of Specialised Translation*, vol. 11, no. 2009, pp. 31–52, 2009.

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

[5] B. A. Lively and S. Pressey, "A method for measuring the" vocabulary burden" of textbooks: Educational administration and supervision," *A method for measuring the" vocabulary burden" of textbooks: Educational Administration and Supervision*, 1923.

[6] E. Dale, "A comparison of two word lists," *Educational Research Bulletin*, pp. 484–489, 1931.

[7] 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

[8] W. A. McCall and L. M. Crabbs, *Standard Test Lessons in Reading...* Teachers College, Columbia University, Bureau of Publications, 1925, no. 2.

[9] J. N. Farr, J. J. Jenkins, and D. G. Paterson, "Simplification of flesch reading ease formula." *Journal of applied psychology*, vol. 35, no. 5, p. 333, 1951.

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

[11] D. R. McCallum and J. L. Peterson, "Computer-based readability indexes," in *Proceedings of the ACM'82 Conference*, 1982, pp. 44–48.

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

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

[14] G. H. Mc Laughlin, "Smog grading-a new readability formula," *Journal of reading*, vol. 12, no. 8, pp. 639–646, 1969.

[15] E. U. Coke and E. Rothkopf, "Note on a simple algorithm for a computer-produced reading ease score," 1969. [Online]. Available:

https://doi.org/10.1037/e527392009-001

[16] 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

[17] 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

[18] 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.

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

[20] 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.

[21] B. Bhagoliwal, "Readability formulae: Their reliability, validity and applicability in hindi," *Journal of Education and Psychology*, vol. 19, no. 1, 1961.

[22] R. K. Agnihotri and A. L. Khanna, "Evaluating the readability of school textbooks: An indian study," *Journal of Reading*, vol. 35, no. 4, pp. 282–288, 1991.

[23] 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.

[24] 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.

[25] I. Lorge, "Predicting readability." *Teachers college record*, 1944.

[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] P. Bhattacharyya, P. Pande, and L. Lupu, "Hindi wordnet." [Online]. Available: https://catalog.ldc.upenn.edu/LDC2008L02

[28] G. Venugopal-Wairagade, J. R., and D. Pramod, "Novel language resources for hindi: An aesthetics text corpus and a comprehensive stop lemma list," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 1, 2020. [Online]. Available: https://doi.org/10.14569/ijacsa.2020.0110130

[29] L. Ramamoorthy, N. Choudhary, J. K. Singh, Richa, A. Sinha, D. Mishra, A. K. Tripathi, A. Debsharma, S. K. Awasthi, and M. e. a. Pathak, "A gold standard hindi raw text corpus," 2019. [Online]. Available: https://data.ldcil.org/text/text-raw-corpus/a-gold-standard-hindi-raw-text-corpus

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