A Novel Readability Complexity Score for Gujarati Idiomatic Text

Jatin C. Modh[1]
Research Scholar
Gujarat Technological University
Ahmedabad, India

Jatinderkumar R. Saini[2]*
Symbiosis Institute of Computer Studies and Research, Symbiosis International (Deemed University)
Pune, India

Ketan Kotecha[3]
Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International (Deemed University)
Pune, India

***Abstract*—Gujarati language is used for conversation by more than 55 million people worldwide and it is more than 1000 years old language. It is the chief language of the Indian state of Gujarat. There are many dialects of Gujarati like Standard Gujarati, Amdawadi Gujarati, Kathiawadi Gujarati, Kutchi Gujarati etc. The Gujarati language is very rich in morphology like other Indo-Aryan languages like Hindi. Many readability tests are available in the English language, but no readability complexity test is available for the Gujarati idiomatic text. The Complexity score is the sub concept of the readability test. In order to define complexity level of Gujarati text, complexity score of Gujarati text is calculated. We deployed a novel readability complexity score calculation method in which we considered the number of letters of each word, the number of diacritics of each word, Gujarati idiomatic text of n-gram where n=1 to 9, Gujarati idiomatic text of m-meaning idioms where m=1 to 7. The complexity score is calculated as the sum of word complexity score, diacritics complexity score, n-gram complexity score of Gujarati idioms and m-meaning complexity score of Gujarati idioms. We emphasized Gujarati idiomatic text for the calculation of complexity score as idioms make the text more complex to understand. This is an innovative and first of its kind work in the research community of Gujarati language. The results are hopeful enough to employ the suggested complexity score method for developing a readability test method for natural language processing tasks for the Gujarati language.**

Keywords*—*Complexity; Gujarati; idiomatic text; natural language processing (NLP); readability

I. Introduction

Gujarati language is named after the people of Gurjar people who are said to have established in the middle of the 5th century CE. Gujarati language is used by more than 55 million people worldwide and it is more than 1000 years old language based on Indo-Aryan languages. Gujarati language stands in 26th position among the most spoken native language in the world. Gujaratis are spread all over the world. It is the chief language of the Indian state of Gujarat. It is also main language in the union territories of Daman and Diu, Dadra and Nagar Haveli. Outside of India, it is spoken all over the world in many countries like United States, Canada, UK, Southeast African countries etc. There are many dialects of Gujarati like Standard Gujarati, Amdawadi Gujarati, Kathiawadi Gujarati, Kutchi Gujarati etc. The spelling of Gujarati words is based on pronunciation [1][2].

A. Gujarati Script

Gujarati is written similar to the Devanagari script except it does not have the horizontal line above characters. The Gujarati alphabet has mainly 34 consonants, 13 vowels and 10 digits working as a building block of the Gujarati language. Sarth Gujarati dictionary consists more than 65000 words excluding technical or slang words [3]. Gujarat vowels and Gujarati consonants can be written as independent letters or by combining with diacritic marks. Diacritics play a very important role in building meaningful words and thus vocabulary of the Gujarati language. Fig. 1 shows the use of diacritics with the letter **ત.** Gujarati diacritics and conjuncts make Gujarati script more effective for written and communication purposes [4][5].

B. Gujarati idioms

An idiom is a group of words but whose meaning is established by the usage and not as the literal meaning of its separate words. Gujarati people are using Gujarati idioms for expressing thoughts, feelings and messages. Gujarati idioms are not understandable for non-Gujarati people as well as for children of a lower standard. Gujarati idioms can be understood by the surrounding context information [6]. Gujarati idioms can be classified on the base of N-grams and on the base of the number of m-meanings [8]. Gujarati idioms can also be classified as static idioms versus inflected idioms. Here we consider idioms as unfamiliar words. Example of Gujarati idiom is જળ લેવું ‘jala levum’ i.e. to take a vow. It is bigram/2-gram and single-meaning idiom.

C. Text Complexity

English language consists of 26 alphabets with 21 consonants and 5 vowels for writing. Generally, three aspects are used to decide the complexity of the English text: quantitative measures, qualitative measures and concerns involving to the reader and task [7]. The Gujarati language is morphologically very rich compared to the English language. The Gujarati language consists of 18 diacritics [6]. Diacritics make many possible word formations by suffixing or prefixing any letter. Using diacritics various inflectional forms are possible for Gujarati verbs and Gujarati nouns [9]. Here only quantitative measures are considered for complexity as our text is just in written form. Factors such as sentence, word length and the frequency of unfamiliar words are used as quantitative measures of text complexity.

*Corresponding Author

Independent vowels	અ a	આ aa	ઇ i	ઈ ee	ઉ u	ઊ oo	એ e	ઐ ai	ઓ o	ઔ au	અં am	અઃ Ah	ઋ ru
Common Diacritics→		ા	િ	ી	ુ	ૂ	ે	ૈ	ો	ૌ	ં	ઃ	ૃ
ત + Diacritics →		તા	તિ	તી	તુ	તૂ	તે	તૈ	તો	તૌ	તં	તઃ	તૃ

Fig. 1. Use of Diacritics in the Building Gujarati Conjuncts with Letter ત.

The rest of the paper is organized as follows: Section II corresponds to the literature review related to text complexity and Gujarati text; Section III represents the methodology including collection of idioms data and the method of calculating Gujarati text complexity; Section IV covers the results and analysis; finally, the limitations, conclusion and future work are represented in Section V.

II. Related Literature Review

A readability score is computer calculated score which roughly decides what level of knowledge needed by someone to be able to read a text easily. Various researches have been performed for the study of the readability and complexity of the various languages. Various work related to readability formula have been carried out.

Harvey [7] represented three-part model for measuring text complexity namely qualitative measures, quantitative measures and reader & task. Quantitative measures consider more lexile level text as more complex than less lexile text. A qualitative factor considers layout, text structure, language features, purpose and meaning etc descriptors. Reader & task is dependent on the professional judgment of teachers about the complex text. Author used a Rubric - a set of guidelines to decide the complexity of the English text.

Uccelli [10] considered parameters like word length, frequency of unfamiliar terms, sentence length and text cohesion for the quantitative dimension of the complexity of English language text. The author emphasized that multiple themes, multiple perspectives, content-specific knowledge, figurative or ambiguous language make English text very complex text.

Anet [11] defined text complexity as easy or hard text in terms of reading based on qualitative and quantitative text features. Important quantitative parameters for defining text complexity are structure, meaning or purpose, language and knowledge requirement for particular English text.

Barge [12] calculated the English text complexity Rubric using 10 dimensions; each dimension can receive a score between 0 and 10 to indicate the optimal benefit for students. 100 points is the best possible overall score for a text and interpreted collective text scores depend on the different points. The rubric provides a framework to assist educators.

Flesch and Kincaid [13] designed readability tests to indicate the difficulty of English passages to understand. They represented two tests namely Flesch Reading-Ease and Flesch-Kincaid Grade level. Same core measures of sentence length and word length are used by the authors for the two tests.

Tillman and Hagberg [14] used Swedish and English language to test the compatibility of readability algorithms. They tested three algorithms namely Coleman-Liau index (CLI), Lasbarhetsindex (LIX) and Automated Readability Index (ARI) on Wikipedia articles. Authors concluded that CLI seem to perform less well on higher level text but works excellent on the Bible like easy to read text in Swedish and English languages, whereas LIX and ARI work on average as well as hard texts in both Swedish and English languages.

Venugopal et al. [15][16] analyzed the complex words in Hindi language sentences and experimented with whether classical readability parameters of the English language can be applied to the Hindi language or not for determining the complexity of the word. They demonstrated that the frequency parameter plays an important role in determining the complexity of a word in Hindi sentence. As per their study, the length of a word is not a significant factor; the number of syllables plays an important predictor of word complexity. Researchers used five tree-based ensemble models out of a total of eight classifiers to extract the important features.

Sinha et al. [17] presented that the English readability formulas are not helpful for Hindi and Bangla languages. They proposed two new readability models for Hindi text documents and Bangla text documents. They customized standard structural parameters like word length, sentence length, number of syllables/word, number of polysyllabic words, number of consonant-conjuncts and number of polysyllabic words per 30 sentences.

Mehta and Majumder [18] explored large-scale media text of three Indo-Aryan languages Gujarati, Bengali, and Hindi as a part of quantitative analysis. As per their statistical study of the corpus, Bengali piece of writing might be more difficult to read than Hindi or Gujarati; Gujarati corpus has more diversity in vocabulary and it contains double type-token ratio than that of Bengali; Hindi is less artificial compare to Gujarati but more compared to Bengali, etc.

Modh and Saini [19][20] collected 2-gram to 9-gram Gujarati idioms and classified them as single-meaning to seven-meaning idioms based on a number of meanings. Authors [6] detected Gujarati idioms from the entered text using diacritics and suffix-based rules. Researchers [8] also exploited IndoWordNet for deciding the meaning of idioms on the base of surrounding contextual information.

Based on this exhaustive literature assessment and evaluation, English language text is analyzed by many researchers in detail for deciding the readability score of the English text by applying different standard parameters. Indo-Aryan languages like Hindi, Bengali and Gujarati are analyzed by some researchers by comparing it with English parameters. Very less work is done specially for Gujarati language text. No researchers have calculated the readability complexity score of

the Gujarati idiomatic text and No other researchers have tried to identify Gujarati idioms from the Gujarati text.

The paper highlights on the study of the complexity of Gujarati text by considering parameters like the number of letters in the individual word and the number of diacritics of the individual word. This paper also considers the presence of idioms in the text and also considers the type of idioms in the text and decides the complexity level of the Gujarati text. The extent of this paper is to analyze letters, diacritics, words and idioms within Gujarati text. This deployment helps in the study of the complexity of Gujarati idiomatic text.

III. Methodology

For the calculation of the complexity score of Gujarati text, four parameters are considered (1) the number of letters of each word (2) the number of diacritics of each word (3) the number of Gujarati idioms. If Gujarati idioms are found in the text, then the idiom(s) are classified in two ways: N-gram classification and M-meaning classification. Different complexity points are allocated to different classifications of idioms. The complexity score is calculated as the summation of meaning complexity, gram complexity, word complexity and diacritics complexity.

Complexity Score=Meaning Complexity Score + Gram Complexity Score + Word Complexity Score + Diacritics Complexity Score

A. Collection of Data

By and large 3472 distinct Gujarati idioms are accumulated from different Gujarati language resources [21][22]. Idiom data collection is basically for the recognition of Gujarati idioms from the Gujarati text.

B. N-Gram Idiom Classification and Complexity Points

Idioms are classified on the basis of N-gram model. Idioms can be classified as 2-gram or bigram, trigram or 3-gram, 4-gram or four-gram, 5-gram, 6-gram, 7-gram, 8-gram, 9-gram. Idiom up to 9-gram was found. 1-gram idioms are specific personage idioms that represent the historical or fictional special character identity in a play. Example of 7-gram Gujarati idiom is **રાન રાન ને પાન પાન થઈ જવું** 'rana rana ne pana pana thai javum' i.e. getting into a bad situation.

Table I shows the classification of idioms on the base of N-grams and their corresponding complexity point calculation method. Bigrams and trigrams are more in number, so both are getting relatively more complexity points compared to other N-gram idioms.

C. M-Meaning Idiom Classification and Complexity Points

Idioms are also classified on the base of their meanings. Gujarati Idiom has a single meaning or more than one meaning. For single meaning idioms, a dictionary based approach is used to understand the meaning of an idiom, but for multiple meaning idioms, surrounding contextual information is needed to understand the idiomatic text. So it is complex to understand multiple-meaning idioms. So M-meaning idioms, corresponding M-complexity points are assigned. Table II shows the classification of M-meaning idioms and corresponding complexity points for the calculation of the complexity score. Gujarati Idioms are found from single meaning to seven meaning idioms. More complexity points are assigned for 7-meaning idioms as it requires more effort to understand by studying the surrounding contextual text.

For example **ઠેકાણું કરવું** 'thekanum karavum' is a 7-meaning idiom as it has 7 different possible meanings depending upon the context like **ઉપયોગમાં લેવું** 'upayogamam levum' i.e. to use, **કન્યાને સારે ઘેર પરણાવવી** 'kanyane sare ghera paranavavi' i.e. marry the bride to the right person, **કાસળ કાઢવું** 'kasala kadhavum' i.e. to kill, **ખલાસ કરવું** 'khalasa karavum' i.e. use-up, **છેવટની ક્રિયા કરવી** 'chevatani kriya karavi' i.e. take the last action, **મારીને દાટી દેવું** 'marine dati devum' i.e. kill and bury, **યોગ્ય સ્થાને ગોઠવી દેવું** 'yogya sthane gothavi devum' i.e. arrange in the right place.

TABLE I. Complexity Point Calculation for each N-Gram Idiom

Sr. No.	N-gram Idioms	Count	(Count/Total Idioms) *10	Complexity Point (Roundup to 2 decimal)
1	Unigrams	58	0.167050691	0.17
2	Bigrams	2102	6.054147465	6.06
3	Trigrams	992	2.857142857	2.86
4	4-Grams	244	0.702764977	0.71
5	5-Grams	63	0.181451613	0.19
6	6-Grams	9	0.025921659	0.03
7	7-grams	2	0.005760369	0.01
8	8-grams	1	0.002880184	0.01
9	9-grams	1	0.002880184	0.01
	Total Idioms	3472		

TABLE II. COMPLEXITY POINT TABLE FOR M-MEANING IDIOMS

Sr. No.	M-meaning idioms	Count	Number of meaning(s)	Complexity Point
1	single-meaning	1806	1	1
2	2-meanings	953	2	2
3	3-meanings	504	3	3
4	4-meanings	193	4	4
5	5-meanings	13	5	5
6	6-meanings	1	6	6
7	7-meanings	2	7	7
	Total Idioms	3472		

D. Diacritics Complexity Score

If there are no diacritics in the Gujarati word, then the particular word is considered simple and easy to read. For example, Gujarati word **રમઝમ** 'ramzam' i.e. ramzam has no diacritics. Another example of a Gujarati word, **ચાદર** 'chadar' i.e. sheet has 1 diacritics. If there are more diacritics in the particular word, then the particular word is difficult to read. If the count of diacritics of a particular word is 0 or 1, then that particular word is considered as simple, so 0 complexity point is assigned. If the count of diacritics of a particular word is 2, then 0.2 complexity point is assigned. If the count of diacritics of a particular word is 3 or 4, then 0.5 complexity point is assigned. If the count of diacritics of a particular word is 5 or 6, then 1 complexity point is assigned. If the count of diacritics of a particular word is greater than or equal to 7, then 2 complexity point is assigned. Table III shows the complexity point table on the base of number of diacritics of a particular word.

E. Word Complexity Score

If the count of letters of a particular word is 1, 2 or 3, then that word is considered as simple, so 0 complexity point is assigned. If the count of letters of a particular word is 4 or 5, then 0.5 complexity point is assigned. If the count of letters of a particular word is 6 or 7, then 1 complexity point is assigned. If the count of letters of a particular word is greater than or equal to 8, then a 2 complexity point is assigned. Table IV shows the complexity point table on the base of the number of letters of a particular word.

F. Database of Idioms

An Idiom database is required to store the collected Gujarati idioms. This idiom database is used to identify idioms from the input text to decide the complexity of the Gujarati idiomatic text. Idiom column stores the base form of the idiom in the idiom database. Fields like idiom, Gujarati meaning of idiom, English meaning of idiom and other related fields are created as a part of the Idiom database [6][23].

TABLE III. COMPLEXITY POINT TABLE ON THE BASE OF NUMBER OF DIACRITICS OF PARTICULAR WORD

Sr. No.	No. of diacritics of particular word	Complexity Point	Example
1	0	0	રમઝમ 'ramzam' i.e. ramzam
2	1	0	ચાદર 'chadar' i.e. sheet
3	2	0.2	વાદળી 'vadali' i.e. blue
4	3 to 4	0.5	ચાદરમાં 'chadarman' i.e. in the sheet
5	5 to 6	1	ચીડિયાપણું 'chidiyapanum' i.e. irritability
6	Greater than or equal to 7	2	પ્રતિદ્વંદ્વિતા 'pratidhvandhita' i.e. competition

TABLE IV. COMPLEXITY POINT TABLE ON THE BASE OF NUMBER OF LETTERS OF PARTICULAR WORD

Sr. No.	Number of letters of particular word	Complexity Point	Example
1	1 to 3	0	આકાશ 'aakash' i.e sky
2	4 to 5	0.5	બતાવવી 'batavavi' i.e. showing
3	6 to 7	1	પ્રયોજનભૂત 'prayojanbhut' i.e. purposeful
4	Greater than or equal to 8	2	તત્વજ્ઞાનીઓનો 'tatvagnaniono' i.e. of philosophers

G. Proposed Model

Fig. 2 explains the steps for the proposed algorithm/model.

Step 1: Accept the Gujarati text from the user.
Step 2: Pre-processing step
2.1: Eliminate whitespaces from starting and ending side of the text
2.2: Eliminate all whitespaces in between the text
Step 3: Tokenize all the words of entered text.
Step 4: Eliminate Gujarati stop words from the entered text.
Step 5: Find out Gujarati idioms from the entered text using the idiom database
Step 6: Calculate the gram-complexity score for idioms as per Table I.
Step 7: Calculate the meaning-complexity score for idioms as per Table II.
Step 8: Count the number of letters of individual word
Step 9: Count the number of diacritics of individual word
Step 10: Calculate diacritics complexity score as per Table III.
Step 11: Calculate word complexity score as per Table IV.
Step 12: Calculate complexity score=Gram-complexity score + Meaning-complexity score + Diacritics complexity score + Word complexity score
Step 13: Display complexity level results of Input text.

Fig. 2. Algorithm for the Proposed Model.

The entered input is the Gujarati text which may or may not contain any unfamiliar words, including the Gujarati idioms. The output will be the analysis of Gujarati text with complexity score, which takes into consideration various factors, and the corresponding complexity level.

IV. Result and Analysis

Gujarati text containing zero or more idioms is given as an input and output shows the related complexity score and complexity level of the inputted Gujarati text. The algorithm ignores the stop words in calculating complexity scores. Output also shows the stop words found in the input text. It also displays total words, total stop words, total letters, and total diacritics used in the input Gujarati text. It calculates Gram complexity score, meaning complexity score, diacritics complexity score and word complexity score as per weight defined in Table I, Table II, Table III and Table IV. The proposed model implements Table V for showing the complexity type or complexity level as an output.

We now present a few examples for the execution of the proposed algorithm for calculating the novel complexity score for the different instances of the Gujarati text. In Example 1, Example 2 and Example 3, different Gujarati text is given as an input. In Example 1, the input text is taken from the standard 1 Gujarati textbook. The output confirms that the complexity type of the text is SIMPLE. This is expected for the text used for teaching the first graders in the age group of generally 5 to 6 years.

TABLE V. Complexity Score Interpretation Table

Sr. No.	Complexity Score	Complexity Type	Notes
1	0.0-20.0	SIMPLE	Very easy words.
2	20.0-40.0	FAIRLY SIMPLE	Fairly Easy.
3	40.0-60.0	MEDIUM	Medium complexity
4	60.0-80.0	COMPLEX	Complex
5	80.0 or more	VERY COMPLEX	Extremely complex

Example1:

INPUT TEXT=વરસાદ આવે રમઝમ વાદળી ચાલે ઝમઝમ, મોટા મોટા છાંટા પડતા આભથી એ નીચે સરતા. આકાશ ગાજે ધમધમ વીજ ચમકની ચમચમ, ધરતી લીલી ચાદર ઓઢે લીલી ચાદરમાં ધરતી પોઢે. ઢમઢમ ઢોલ વગડાવો, વરસાદને સૌ વધાવો.
'varasada ave ramajhama vadali cale jhamajhama, mota mota chanta padata abhathi e nice sarata. akasa gaje dhamadhama vija camakati camacama, dharati lili cadara odhe lili cadaramam dharati podhe. dhamadhama dhola vagadavo, varasadane sau vadhavo.'

OUTPUT:
STOP WORDS FOUND----> આવે, એ, નીચે,
'ave, e, nice,'
Total Words in input Text: 34
Total Idioms Found: 0
Meaning Complexity Score: 0
Gram Complexity Score: 0
Word Complexity Score: 6
Diacritics Complexity Score: 3.2
Total letters in input: 97
Total diacritics in input: 41
Total stop words in input: 3
Complexity Score = 9.2
Complexity Type = **SIMPLE**

In Example 2, the input text contains the collection of 13 idioms. Output identifies these 13 idioms and from these 13 idioms, 8 idioms are with 1-meaning, 3 idioms are with 2 meanings, 1 idiom with 3 meanings and 1 idiom with 4 meanings. Output also identifies different N-gram wise idioms. Corresponding meaning complexity score and gram complexity score are calculated. Word complexity score and Diacritics complexity score is also calculated. Finally, the complexity score is calculated and the complexity type is decided on the base of the range of complexity score.

Example2:

INPUT TEXT=એક કાને સાંભળી બીજે કાને કાઢી નાખવું ઢ સંસાર માંડવો આગ લાગવી અક્કલ ચરવા જવી આંખમાં પાણી આવવું આકાશ પાતાળ જેટલું અંતર આંખ બતાવવી અક્કડ ને અક્કડ રહેવું જમીન પર પગ ન મૂકવો નાક ઉપર માખી ન બેસવા દેવી એકે પથ્થર ઉથામ્યા વગરનો ન રહેવો રાત કહે તો રાત દહાડો કહે તો દહાડો
'eka kane sambhali bije kane kadhi nakhavum dha sansara mandavo aga lagavi akkala carava javi ankhamam pani avavum akasa patala jetalum antara ankha batavavi akkada ne akkada rahevum jamina para paga na mukavo naka upara makhi na besava devi eke paththara uthamya vagarano na rahevo rata kahe to rata dahado kahe to dahado'

OUTPUT:
STOP WORDS FOUND----> એક, જેટલું, ને, રહેવું, પર, ન, ઉપર, ન, ન, તો, તો,
'eka, jetalum, ne, rahevum, para, na, upara, na, na, to, to,'

Total Words in input Text: 53
Total Idioms Found: 13

8 Idioms With 1 Meaning(s)
3 Idioms With 2 Meaning(s)
1 Idioms With 3 Meaning(s)
1 Idioms With 4 Meaning(s)
Meaning Complexity Score: 21

1 Idioms With 8 Gram(s)
1 Idioms With 7 Gram(s)
2 Idioms With 6 Gram(s)
1 Idioms With 5 Gram(s)
2 Idioms With 4 Gram(s)
2 Idioms With 3 Gram(s)
3 Idioms With 2 Gram(s)

1 Idioms With 1 Gram(s)
Gram Complexity Score: 26.5

Word Complexity Score: 3.5
Diacritics Complexity Score: 5.9

Total letters in input: 114
Total diacritics in input: 66
Total stop words in input: 11

Complexity Score = 56.9
Complexity Type = **MEDIUM**

In Example 3, the complexity score is calculated as 75.3, which is in the range of 60.0-80.0, so the output of the complexity type is COMPLEX.

Example3:

INPUT TEXT=અંતરવેદના અંધાધૂંધી અતિસૌરભ હાથ ઝાલ અનુસંધાન અવસન્નતા અવસન્નત્વ આજ્ઞાંકિત આદીનવ આમ્રવૃક્ષ ઇંદ્રશસ્ત્ર ઇંદ્રાયુધ ઇમ્તિહાન ઉપદ્રવ ત્રિદશાંકુશ ત્રિદશાયુધ ધાતુરાજક નિરીક્ષણ પરિચારક પરેશાની પર્યેષણા પિકવલ્લભ પૂછપરછ પ્રતિકુળ પ્રિયાંબુ ભોગવિલાસ અંતર રાખ મજ્જારસ મનોવ્યથા મુશ્કેલી મેઘજ્યોતિ મેઘભૂતિ રતિકલહ રતિકેલિ રતિસંહતિ રતિસુખ વજ્રાશનિ વસંતદૂત વસંતદ્રુ વસંતદ્રુમ વિટંબણા વિરુદ્ધતા વિષયભોગ વિષયસુખ વ્યાકુલપણું વ્યાકુળતા શતકોટી સહાયરૂપ સૌદામની સૌદામિની સ્ત્રીગમન સ્ત્રીસંસર્ગ સ્ત્રીસુખ સ્ત્રીસેવન હેરાનગત
'antaravedana andhadhundhi atisaurabha hatha jhala anusandhana avasannata avasannatva ajnankita adinava amravrksa indrasastra indrayudha imtihana upadrava tridasankusa tridasayudha dhaturajaka niriksana paricaraka paresani paryesana pikavallabha puchaparacha pratikula priyambu bhogavilasa antara rakha majjarasa manovyatha muskeli meghajyoti meghabhuti ratikalaha ratikeli ratisanhati ratisukha vajrasani vasantaduta vasantadru vasantadruma vitambana virud'dhata visayabhoga visayasukha vyakulapanum vyakulata satakoti sahayarupa saudamani saudamini strigamana strisansarga strisukha strisevana heranagata'

OUTPUT:
STOP WORDS FOUND---->
Total Words in input Text: 57
Total Idioms Found: 2

1 Idioms With 2 Meaning(s)
1 Idioms With 5 Meaning(s)
Meaning Complexity Score: 7

2 Idioms With 2 Gram(s)
Gram Complexity Score: 12.2

Word Complexity Score: 33
Diacritics Complexity Score: 23.1

Total letters in input: 278
Total diacritics in input: 163
Total stop words in input: 0

Complexity Score = 75.3
Complexity Type = **COMPLEX**

V. Conclusion, Limitations and Future Work

The proposed Gujarati text complexity prediction model was successfully implemented and it was based on the number of diacritics of the individual word, the number of letters of the individual word and on the number of idioms. Different complexity points are considered on the basis of N-gram idioms and M-meaning idioms. Gujarati idioms are considered as unfamiliar words to understand the Gujarati text. The complexity score of Gujarati text is calculated as the summation of diacritics complexity points, word complexity points, N-gram idiom complexity points and M-meaning idiom complexity points.

The proposed model could not recognize idioms those are not stored in the idiom database for assigning complexity points. Future work is to assemble all Gujarati idioms to correct this drawback. In the future enhancement of the model, particular domain vocabulary can be used for defining complexity levels.

Based on the outcome achieved, it is advocated that the projected readability complexity score calculation method is worth implementing in the real world for the community of Gujarati language. To the best of our knowledge, it is the first and novel readability complexity score calculation method and complexity type prediction method for the Gujarati Idiomatic text. The proposed method considers the Gujarati idioms as unfamiliar words and assigns weightage accordingly by dynamically detecting them from the input text. The proposed method opens the path for other Gujarati language researchers in defining readability levels for Gujarati text as well as natural language processing tasks for the Gujarati language.

References

[1] Wikipedia, "Gujarati language", https://en.wikipedia.org/wiki/Gujarati_language (accessed March 23, 2022).

[2] Yourdictionary, "Gujarati Language Overview and Common Words"; Available Online: https://reference.yourdictionary.com/other-languages/gujarati-language-words.html (accessed March 23, 2022).

[3] GujaratiLexicon, "Let's Learn Gujarati"; Available Online: http://www.letslearngujarati.com/vowels (accessed March 23, 2022).

[4] Audichya M. and Saini J.R., 2019, "An Overview of Optical Character Recognition for Gujarati Typed and Handwritten Characters", Available online: https://www.researchgate.net/publication/350173104.

[5] Rakholia R.M. and Saini J.R., 2015, "The Design and Implementation of Diacritic Extraction Technique for Gujarati Written Script using Unicode Transformation Format", proc. of IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT-2015), Coimbatore, India, vol. 2, pages 654-659, Available online: https://ieeexplore.ieee.org/document/7226037.

[6] Modh J.C. and Saini J.R., "Dynamic Phrase Generation for Detection of Idioms of Gujarati Language using Diacritics and Suffix-based Rules", International Journal of Advanced Computer Science and Applications (IJACSA), 12(7), 2021. http://dx.doi.org/10.14569/IJACSA.2021.0120728.

[7] Harvey S., "A Beginner's Guide to Text Complexity", Generation Ready, 2013; Available online: https://www.generationready.com/wp-content/uploads/2021/04/Beginners-Guide-to-Text-Complexity.pdf.

[8] Modh J.C. and Saini J.R., "Using IndoWordNet for Contextually Improved Machine Translation of Gujarati Idioms", International Journal of Advanced Computer Science and Applications (IJACSA), 12(1), 2021. http://dx.doi.org/10.14569/IJACSA.2021.0120128.

[9] Kapadia U. and Desai A., "Rule Based Gujarati Morphological Analyzer", IJCSI International Journal of Computer Science Issues, Volume 14, Issue 2, March 2017, https://www.ijcsi.org/papers/IJCSI-14-2-30-35.pdf.

[10] Uccelli P., "Why do so many adolescents struggle with content-area reading?", Available Online: https://iris.peabody.vanderbilt.edu/module/sec-rdng2/cresource/q1/p02/ (accessed March 23, 2022).

[11] The Achievement Network Ltd, "Text Complexity", Available Online: https://www.achievementnetwork.org/anetblog/eduspeak/text-complexity (accessed March 23, 2022).

[12] Barge J., "Common Core Georgia Performance Standards Text Complexity Rubric", Georgia Department of Education, 2011; Available

online: https://www.gpb.org/sites/default/files/2020-06/handout-1-ccgps-ela-textcomplexity-guide.pdf.

[13] Wikipedia, "Flesch–Kincaid readability tests", Available Online: https://en.wikipedia.org/wiki/Flesch–Kincaid_readability_tests (accessed March 23, 2022).

[14] Tillman R. and Hagberg L. (2014), "Readability algorithms compability on multiple languages", Digitala Vetenskapliga Arkivet (DiVA), Stockholm, Available Online: https://www.diva-portal.org/smash/get/diva2:721646/FULLTEXT01.pdf.

[15] Venugopal G., Dhanya P., Saini J.R. (2021), "Analyzing Complex Words in Hindi using Parameters of Classical Readability Formulae (Part 1)", Computer Science Journal of Moldova, 29(3):366-387. Online: http://www.math.md/files/csjm/v29-n3/v29-n3-(pp366-387).pdf.

[16] Venugopal G., Dhanya P., Saini J.R. (2022), "Revisiting the Role of Classical Readability Formulae Parameters in Complex Word Identification (Part 2)", Computer Science Journal of Moldova, 30(1):49-63. Available Online: http://www.math.md/files/csjm/v30-n1/v30-n1-(pp49-63).pdf.

[17] Sinha M., Sharma S., Dasgupta T. and Basu A. (2012), "New Readability Measures for Bangla and Hindi Texts", Proceedings of COLING 2012: Posters, pages 1141–1150. Available Online: https://aclanthology.org/C12-2111.pdf.

[18] Mehta P. and Majumder P. (2014), "Large Scale Quantitative Analysis of three Indo-Aryan Languages", Journal of Quantitative Linguistics, Available Online: https://www.researchgate.net/publication/272496714_Large_Scale_Quantitative_Analysis_of_three_Indo-Aryan_Languages.

[19] Modh J.C. and Saini J.R., 2018, "A Study of Machine Translation Approaches for Gujarati Language", International Journal of Advanced Research in Computer Science, Volume 9, No. 1, January-February 2018, pages 285-288; Available online: ijarcs.info/index.php/Ijarcs/article/download/5266/4497.

[20] Modh J.C. and Saini J.R., "Context Based MTS for Translating Gujarati Trigram and Bigram Idioms to English," 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1-6, doi: 10.1109/INCET49848.2020.9154112.

[21] GujaratiLexicon, Gujaratilexicon.com, Available online: http://www.letslearngujarati.com/about-us (accessed March 23, 2022).

[22] Rudhiprayog ane kahevatsangrah, published by Director of Languages, Gujarat State, Gandhinagar. 2010.

[23] Saini J.R. and Modh J.C., "GIdTra: A dictionary-based MTS for translating Gujarati bigram idioms to English," 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), 2016, pp. 192-196, doi: 10.1109/PDGC.2016.7913143.