Automated Lexical Simplification of Hindi Text

Thesis Submitted to Symbiosis International (Deemed University)

For award of the degree of

DOCTOR OF PHILOSOPHY

Faculty of Computer Studies

GAYATRI VENUGOPAL TANDON
(PRN: 16039001002)

Under the Guidance of

Dr. Dhanya Pramod
Professor and Director, Symbiosis Centre for Information Technology
and
Dean, Faculty of Computer Studies, SIU, Pune

SYMBIOSIS INTERNATIONAL (DEEMED UNIVERSITY), Pune – 412115

2024

THESIS CERTIFICATE

1. The Thesis entitled **Automated Lexical Simplification of Hindi Text** submitted to the Symbiosis International (Deemed University), Pune for the award of Ph.D. Degree under the Faculty of Computer Studies is based on my original work carried out under the guidance of Dr. Dhanya Pramod from September 2016 to February 2024.

2. The Research Work has not been submitted elsewhere for award of any degree.

3. The material borrowed from other source and incorporated in Thesis has been duly acknowledge and/or referenced.

4. I understand that I would be held responsible and accountable for plagiarism, if any, detected later on.

5. Research papers published based on the research conducted out of and in the course of the study leading to Ph.D. are duly credited to SIU and appended to the thesis and has not formed the basis for the award of any degree, diploma, associateship, fellowship, titles in this or any other University or other institution of Higher learning.

Date: ……

Signature of Research Scholar
Gayatri Venugopal Tandon
PRN: 16039001002

Counter signed by Research Supervisor
Dr. Dhanya Pramod
Professor and Director, Symbiosis Centre for Information Technology
and
Dean, Faculty of Computer Studies, SIU, Pune

Acknowledgements

I extend my heartfelt gratitude to the individuals who have been pivotal in the realization of this doctoral thesis, a journey that spanned an extended period of time. This work would not have been possible without the unwavering support, patience, and understanding of numerous individuals.

First and foremost, I am deeply indebted to my esteemed advisor, Professor Dr. Dhanya Pramod, for her empathy, patience, steadfast guidance, invaluable insights, and continuous encouragement throughout the extended course of this research. Dr. Pramod's expertise and unwavering commitment have been instrumental in shaping the direction and quality of this thesis.

I wish to express my sincere appreciation to the members of the Research Recognition Committee, the Research Advisory Committee, the Minor Research Project panel, and the Independent Ethics Committee of Symbiosis International University, for their perceptive critiques, constructive feedback, and for sponsoring the research project, respectively.

I am profoundly grateful to Symbiosis International (Deemed University) for granting necessary extensions, recognizing the challenges and unforeseen obstacles that contributed to the prolonged duration of this research, and also for sponsoring part of my work under the Minor Research Projects programme. The university's support allowed for the thorough exploration and comprehensive analysis that underpins this work.

Throughout this journey, I encountered numerous personal and professional hardships and obstacles that tested my perseverance. The encouragement, camaraderie, and shared experiences within the academic community, particularly among my colleagues and peers, provided the resilience needed to surmount these challenges. To name a few mentors and colleagues who helped me in this journey, I would like to mention Dr. Urvashi Rathod, Dr. Haridas Acharya, Dr. Jatinderkumar R. Saini, Dr. Lalit Kathpalia, Dr. Naganathan Rengasari, Dr. Sachin Naik, Dr. Shilpa Majumdar, Dr. Prafulla Bafna, and Ms. Janhavi Pednekar for their valuable inputs at various stages of this research. This research has benefitted significantly from the inputs of other researchers working in this area, namely, Dr. Ravi Shekhar, Dr. Matthew Shardlow, Dr. Fernando Alva-Manchego, and Dr. Horacio Saggion. I shall always be indebted to the exposure that I gained from their work, feedback, and support.

As a result of entering this field of text simplification, I was able to attend the Association for Computational Linguistics (ACL) in 2020 virtually, and interact with other researchers working in the area of natural language processing. Subsequently, I also got the opportunity to be a member of the programme committee of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022), and contributed in reviewing papers on text simplification.

My parents were my pillars of support throughout this journey. My heartfelt gratitude and thanks are extended to my family, families that supported me during this tenure, and friends, whose unwavering belief in my abilities, their patience, strength and sacrifices, and unending support sustained me through the ups and downs of this seemingly impossible and arduous endeavor. Their presence and words served as a constant source of motivation, and this thesis would not have been complete without the light they bring to my life.

Lastly, I wish to acknowledge and appreciate the dedication of all research participants who generously contributed their time and insights, enabling the comprehensive investigation that forms the foundation of this study.

Executive Summary

The need for lexical simplification arises from the necessity for making content and products accessible to a wide range of individuals, regardless of their abilities. In today's context, accessible technology holds greater significance in fostering inclusivity. Beyond improving the lives of those with impairments, accessible content benefits both, people with impairments and people without impairments. Accessibility pertains to both the tools the researcher use and the content the researcher consume, with a significant aspect being comprehension. Failure to comprehend content can lead to financial losses, accidents, dissatisfaction, and disinterest. Clear instructions are extremely crucial, as their absence has led to accidents. Edward Lee Thorndike noted that readers lose interest when they do not understand the meaning of a word. Legal documents often exemplify this issue. Therefore, plain language, involving simple, comprehensible language, is key for effective communication.

The Plain Language Movement, which began in the 1970s in the United States, aimed to make regulations understandable. This movement spread to various fields and languages. Researchers worldwide strive to make text understandable. It is vital to simplify publicly accessible content for a diverse audience. With the evolving field of natural language processing, the field of text simplification has evolved significantly over the years. Textual content can be simplified in various ways, such as lexical simplification, which substitutes complex words with simpler synonyms, while retaining the context. Lexical simplification is invaluable for people who are unfamiliar with a language, persons with language difficulties, non-native speakers, individuals with low literacy, and children. The researcher explored lexical simplification across languages, noting the need for a bibliometric study to gauge the quantity of work, its diversity, and collaborative efforts made in this field.

Despite the widespread use of Hindi, lexical simplification efforts in the language are limited. The stages of lexical simplification include complex word identification, synonym generation, word sense disambiguation, and synonym ranking. The research aims to identify complex word features, employ a knowledge-based approach for word sense disambiguation, and generate simpler synonyms for the target complex word.

The researcher created a corpus for dataset development, analyzed readability formulae and word lists for relevant features, and incorporated word embeddings. Word sense disambiguation was performed using sense embeddings and human annotation.

This study’s contributions encompass insights into the lexical simplification needs of native and non-native Hindi speakers, an aesthetics corpus, a stop lemma list, a dataset with lexical features, and innovative approaches to complex word identification, dataset creation, and word sense disambiguation.

The research identifies the effectiveness of stop lemmas, presents a publicly available aesthetics corpus, and emphasizes the importance of frequency, syllables, and hyponyms in determining word complexity. The researcher developed a complex word identification model, combining linguistic knowledge and machine learning techniques. The findings underscore the need for delving deeper into personalised lexical simplification, utilising sense embeddings for Hindi word sense disambiguation, and devising readability metrics tailored to Hindi.

This study contributes to enhancing language accessibility, communication efficiency, and NLP applications for Hindi speakers. Extensions of this research could include exploring alternative complex word labeling strategies and model fine-tuning.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ACL	:	Association for Computational Linguistics
AUC	:	Area under the ROC Curve
BERT	:	Bidirectional Encoder Representations from Transformers
CFILT	:	Centre for Indian Language Technology
CNN	:	Convolutional Neural Networks
IoT	:	Internet of Things
MeitY	:	Ministry of Electronics and Information Technology
NLP	:	Natural Language Processing
NLTK	:	Natural Language ToolKit
NMT	:	Neural Machine Translation
ROC	:	Receiver Operating Characteristic
SD	:	Standard Deviation

Chapter 1

Introduction

1.1 Need for Lexical Simplification

Accessibility is an attribute of a product or content that indicates whether everyone can use it regardless of their abilities. Today, the importance of accessible technology has multiplied in order to foster an inclusive society. In addition to enhancing the quality of life for people with disabilities, accessibility also benefits people without disabilities (Henry, 2006). Both, the tangible objects that people use daily, as well as the content that readers consume, can be linked to accessibility. One crucial aspect of writing that is frequently overlooked by content producers is comprehension. Failure to understand the text's meaning may result in loss of money and, regrettably, life. It can also cause dissatisfaction and a lack of interest. For instance, it has been observed that a lack of clear instructions in safety manuals has contributed to traffic accidents (Temnikova, 2012). According to Edward Lee Thorndike (1874–1949), the psychologist who developed the connectionism theory, a reader may lose interest in a text if they are unable to understand the meaning of a term in it (Joncich, 1968). This is evident in legal documents where the phrases are difficult for a layman to understand. The use of basic, comprehensible language to communicate ideas is known as 'plain language'. The term 'plain language' refers to a style of writing that makes it simple for a reader to locate, comprehend, and utilise information from the text. The textual content can be made accessible by being made simpler to read and comprehend.

The United States witnessed a movement named the Plain Language Movement in the 1970s, wherein the objective was to ensure that the regulations should be written in a language that the concerned persons can understand (Locke, 2003). This movement later percolated to different fields such as the food industry, science, and medical journals. Today, numerous researchers across the world are attempting to devise methods to make textual content easy to understand. Despite the fact that plain language has not advanced much in India, it is crucial to make publicly accessible texts simpler so that not only regular readers, but also non-native readers, readers with low literacy, and readers who have reading difficulties like dyslexia, can benefit

from them. The rapidly evolving field of natural language processing encompasses diverse problems including machine translation, sentiment analysis, and named entity recognition, to mention a few. Text simplification is the process of making modifications to a text in such a way that the reader can better understand it without losing any information (Siddharthan, 2014). Text simplification encourages the use of plain language in writings from a variety of fields, including law, education, business, and others. Text can be simplified in a variety of ways, including conceptually, elaboratively, syntactically, and lexically (Paetzold and Specia, 2016a). Lexical simplification is the process of substituting one or more complex words, that is, difficult to understand words, in a specified text with their simpler synonyms, while considering the context of the target complex word (Paetzold and Specia, 2017). Syntactic simplification is the process of changing the syntax, that is, the grammar of a sentence, to make it simpler for the reader to understand (De Belder and Moens, 2010). Here, the word 'complex' indicates a word that is not easy to comprehend by a reader. The researcher does not refer to word complexity in terms of morphological structure.

Lexical simplification has been shown to be beneficial for readers who are unfamiliar with the language, readers who have language-related difficulties (De Belder and Moens, 2010; Rello et al., 2013; Caplan, 1992; Carroll et al., 1999), non-native speakers (Carroll et al., 1998), readers with cognitive impairment (Alarcon et al., 2024), readers who have low literacy levels (Paetzold and Specia, 2016a), and children (De Belder and Moens, 2010). A part of the researcher's study attempts to analyse the lexical simplification work that has been done in different languages around the world. The author observed that although there were studies that extensively reported the nature and details of the work done in this area, there was a need to conduct a bibliometric study to analyse the quantity and diversity of work published in this area over the years, and also to report the key contributors, including authors and countries, and the nature of collaboration in this field.

Today, numerous documents are available in Hindi, the official language of the Indian government. Despite having the third-highest number of speakers worldwide, behind Mandarin Chinese and English, according to Ethnologue[1], there is no study on lexical simplification in Hindi. Furthermore, 43.63 percent of the population speaks Hindi, according to the most recent census conducted in India[2].

[1] https://www.ethnologue.com/guides/ethnologue200

[2] https://censusindia.gov.in/2011census/Language-2011/Statement-4.pdf

Hindi is the official language of the Indian government[1]. It is crucial to create and share content that everyone can understand, as the vocabulary of an individual differs according to their exposure to, and experience with the language. Therefore, it is necessary to make them accessible to readers with different lexical restrictions. This indicates a need to adopt automatic text simplification technologies, as manual content simplification could be time-consuming and labor-intensive. The goal of this research was to simplify words in Hindi text.

In order to take stock of the amount of work done on lexical simplification, the researcher studied the publications from Scopus[2], Web of Science[3], and the Association for Computational Linguistics (ACL) Anthology[4]. Based on the metadata that was available, the researcher observed that majority of the studies focused on English, with a few studies focusing on French, German, Spanish, and more. However, there was a scarcity of studies that focused on lexical simplification of Hindi text.

1.2 Lexical Simplification Stages and Objectives

Shardlow (2014) illustrated the processes in a lexical simplification pipeline that is composed of complex word identification, substitution generation, word sense disambiguation, and synonym ranking. The description of the terms is as follows:

1. Complex Word Identification - The process of automatically identifying a difficult-to-understand word in the text. This approach looks for words that could be difficult for readers to understand, through the use of algorithms and linguistic analysis. Once these words are found, more linguistic improvements or simplifications can be made.
2. Substitution Generation - The process of generating all the synonyms of the word. This process creates an exhaustive list of possible replacements by using lexical databases, language models, or linguistic algorithms. It offers a variety of choices that could act as less complex substitutes for the detected complex word.

[1] https://rajbhasha.gov.in/en/official-language-resolution-1968

[2] https://www.scopus.com/

[3] https://clarivate.com/webofsciencegroup/solutions/web-of-science/

[4] https://aclanthology.org/

3. Substitution Selection - The process of identifying the correct sense of the word according to its context, and subsequently the appropriate synonyms in the sense. To guarantee precise selection, complex computational models such as word sense disambiguation or human annotation may be used.
4. Substitution Ranking - The process of ranking the synonyms according to the context of the target word and according to its complexity. This process takes into account each substitute's suitability in the given context, making sure that the choices made, meet both the intended simplification and the target semantic fit. While ranking potential replacements, variables including simplicity of comprehension, frequency, and relevancy may be taken into account. The end goal is to provide a ranking of alternatives so that the user may make an informed decision on what to simplify next.

The researchers framed the research objectives framed based on the steps of lexical simplification:

Research Objective 1: To identify the features of a complex word in Hindi

Research Objective 2: To determine a suitable method using the knowledge-based approach for word sense disambiguation

Research Objective 3: To validate the model

1.3 Approach

1.3.1 Corpus Creation and Annotation Study for Dataset Creation

The first objective dealt with the identification of features of a complex word. In order to achieve this, the researcher studied existing resources that could be used for complex word identification. Although lexical simplification has become more prominent in several other languages, including English, Brazilian Portuguese, Spanish, French, and more (Paetzold, 2016; Lee and Yeung, 2018; Hmida et al., 2018; Alu´ısio and Gasperin, 2010), there is a scarcity of resources for simplification in Hindi, for instance, the Simple English Wikipedia (Coster and Kauchak, 2011), lists of simple words (Ogden, 1930), and ranked synonym lexicons (Billami et al., 2018). The researcher could not find a dataset that was annotated to distinguish between simple and complex words in Hindi text.

It was necessary to study how certain attributes affected the lexical complexity of sentences given the differences between languages, for instance, the existence of consonant conjuncts in

Hindi words. As a result, it was crucial to create tools that cater to a specific language and that are specifically designed for Hindi text simplification. Hence, one of the sub-objectives was to create resources required for complex word identification in Hindi text.

A corpus that reflects the target domain or language, is one of the fundamental prerequisites for carrying out any supervised machine learning task in Natural Language Processing (NLP). In order to produce a comprehensive list of stop lemmas using their occurrence count in the corpora, the researcher developed a corpus containing work that can be categorised under the aesthetics domain, and used it in combination with other corpora. The corpus is available for public use under the GNU General Public License. The corpus was developed as a result of the challenges the researcher encountered when looking for and obtaining books, stories, and non-fiction works published by both contemporary authors and authors from India's pre-independence era, or before 1947. The researcher also created a stop lemma list so that the machine does not attempt to simply these words. The comprehensive list of stop lemmas developed as part of this effort was constructed from text from several sources, thus making it appropriate for widespread use. The researcher believes that the corpus and stop lemma list are valuable tools for other NLP tasks as well, such as information retrieval, text summarisation, and more.

The biggest challenge that the researcher encountered was the subjective characteristic of the problem, as the reader's vocabulary and familiarity with the word have a significant role in the complexity or incomprehensibility of a word. With the intent to simulate a real-world situation where readers with different levels of vocabulary skills would consume the content, the researcher designed an annotation task, wherein they made an effort to include readers who were conversant in a variety of languages other than Hindi. However, it should be noted that Hindi speakers whose native language is not Hindi, might be familiar with certain words if they are shared with the Sanskrit language, as Sanskrit words are used in various Indian languages. Consequently, it was possible for a non-native Hindi speaker to be as proficient as, or even more so, in Hindi, as compared to a native Hindi speaker.

1.3.2 Analysing Readability Formulae and Word Lists for Identifying Relevant Features

In order to create a complex word identification-cum-simplification model, the researcher aimed to build a classifier that would be trained on the dataset created using the data acquired from the annotation study. To ensure that the researcher considered relevant lexical features, they studied existing readability formulae. Researchers have developed several readability formulae that identify the grade level or the age of the target reader in order to assist content developers in developing material in accordance with the user base of the text (Flesch,1948; Gunning et al., 1952; Fry, 1968; Mc Laughlin, 1969; Coke & Rothkopf, 1969; Kincaid et al., 1975; Coleman & Liau, 1975; Weir & Ritchie, 2006). The use of these metrics for studying the difficulty of words in Hindi text had not yet been investigated, despite the fact that these metrics were readily available for the English language, and were modified for use in languages besides English (Agnihotri & Khanna, 1991). The researcher could use any of these criteria to evaluate the complexity of a word because they all operate at the text level. Similarly, simple word lists created by researchers in the field focused on the English language, and a certain demographic of intended readers, for instance, children in a particular grade. However, there are no equivalent word lists for Hindi. Therefore, in order to identify similarities among the criteria for selection of words, the researcher aimed to study popular English word lists as well as a few non-English word lists.

The majority of the readability formulae are based on English, despite the fact that they have been developed for many different languages. Along with readability formulae, the researcher also analysed the lexical parameters that were utilised to create word lists that contained terms that were regarded as simple. The researcher studied the aspects of words that could be measured and used as inputs for readability formulae. The researcher concentrated on the most prevalent traits in the research, and considered them as potential features of words in various models of classification. The results of the two methods for identifying the key characteristics clearly demonstrated the significance of word frequency and the insignificance of word length. This outcome contradicts the observation that several readability formulae relied on word length. Additionally, the researcher discovered that a major indicator of a word's difficulty that was not accounted for by any formula, was the number of hyponyms it contains.

1.3.3 Analysing Word Embeddings as Features

The meaning of a word is dependent on the context in which it is used (Harris, 1954). Additionally, according to Harris (1954), words that are used in comparable settings frequently have related meanings. The distributional theory, which led to the evolution of word embeddings, is based on this. Word embeddings are vectors that include the semantics of the words (Wiedemann et al., 2019). They represent the meaning of a word as a vector so that words with related meanings have similar representations (Turian et al., 2010). Word embeddings, also known as distributed representations, were utilised by numerous researchers to designate words as complex (Kuru, 2016; Butnaru & Ionescu, 2018; Wani et al., 2018; Sheang, 2019). However, the application of word embeddings to identify complex words in Hindi text had not been investigated yet. This research encompasses the study of word embeddings as features for identifying complex words.

1.3.4 Word Sense Disambiguation

In order to identify substitutes of a word identified as complex, the researcher performed word sense disambiguation so that the correct sense is determined. This involves distinguishing and selecting the precise meaning or sense in which a word is used within a particular context. Given the absence of extensive publicly available datasets with annotations, crucial for training a model to perform this task, the researcher turned to an alternative method. Initially, the researcher made use of ARES embeddings, which were first presented by Scarlini et al. (2020b). These embeddings provide a representation of the several senses associated with words in multiple languages, specifically designed for the challenges of sense identification. The ARES embeddings are noteworthy for encompassing languages found in the vast lexical collection, BabelNet[1]. The process was extended to include human annotation as a supplement to the automated approach. This included the integration of human annotators' insights and assessments, which gave the word sense disambiguation task a more complex layer of understanding. The researcher wanted to improve the accuracy of the word sense disambiguation process by merging human expertise with the strengths of ARES embeddings.

[1] https://babelnet.org/

1.4 Contributions

The key contributions of the research are as follows:

- Valuable insight into the diverse lexical simplification needs of Indian Hindi speakers.
- An aesthetics corpus containing text from Hindi novels, stories, and biographies.
- A stop words list for Hindi text processing
- A dataset (tested for bias) containing lexical features of Hindi words and their labelled complexity values
- Novel approaches for
    - complex word identification in Hindi text
    - Hindi complex word dataset creation using synset based normalisation
    - Hindi word sense disambiguation using a knowledge based approach involving word embeddings

1.5 Findings

After conducting a series of experiments for dataset creation, stop word list creation, complex word identification, and word sense disambiguation, the researcher reports several diverse and unique findings in this thesis. This study demonstrates the effectiveness of stop lemmas over stop words by evaluating existing stop word lists and word clouds, and comparing the Hindi stop lemma list with the English language stop words found in the NLTK Python package.
The research suggests that stop lemma lists could prove valuable in various NLP tasks. The Hindi stop lemma list containing 311 entries, developed in this study is made publicly accessible for further use, as it has been released under an open source license.

The researcher observed that a corpus exclusive to the aesthetics domain spanning over a hundred years does not exist. Therefore the researcher collated texts, pre-processed them and presented a corpus which is now publicly available for further research. The corpus contains 145,508 unique words and 118,266 unique lemmas. This annotation study revealed various aspects of understanding of the language. The researcher created a dataset as a result of this study that consists of 7,321 words annotated by 100 different annotators. The findings of the annotation study reveal significant variances between user categories, with individuals more familiar with the language performing better in agreement values. The research also highlights the importance of frequency, the syllable count, and the hyponym count in determining word

complexity. However, the study suggests the need for a readability metric specifically designed for Hindi at both the word and sentence levels.

The model that the researcher created is a combination of several tree-based ensemble models and has been evaluated using various evaluation metrics such as AUC scores, F1 scores, and accuracy. In comparison to usual machine learning model performances, the findings might not appear to be exceedingly impressive. However, the researcher argues that making a machine predict a word's complexity, which is highly individualised for the reader, is something the reader must take into account. The findings demonstrate the effectiveness of combining linguistic knowledge with machine learning techniques and sheds light on crucial factors impacting word complexity in Hindi sentences. These contributions have the potential to enhance language accessibility, communication efficiency, and natural language processing applications for Hindi speakers. Future research in this field can further explore alternative strategies for complex word labeling and fine-tuning models through hyperparameter tuning.

1.6 Organisation of the Thesis

In Chapter 2, the researcher reports an analysis of the research on lexical simplification that has been done by scholars worldwide. With a comprehensive approach, the researcher studied a wide range of studies, exploring the nuances of simplification strategies in many languages. This broad approach was chosen with the assumption that the techniques used to find complex words in one language may easily cross linguistic boundaries and promote a comprehensive understanding of the topic. In Chapter 3, the details of the study's research design and implementation are reported. The complex decisions and factors that influenced the study are presented in this chapter. Every aspect of the study's methodology, from the choice of data sources to the algorithmic decisions, is explained. The thesis reaches a critical point in Chapter 4, which presents the results of the researcher's empirical work. This chapter presents the findings from rigorous experiments. The researcher describes an empirical exploration journey where the effects of lexical simplification techniques on various linguistic constructs are examined. The researcher reports a variety of findings through data analysis and interpretation, each of which would add to the evolving understanding of the complexities of lexical simplification. In Chapter 5, the researcher explores the interpretations derived from the study's findings.

The researcher uses Chapter 5 as a reflective canvas to paint a thorough picture of the findings' implications. The implications for language comprehension, communication, and educational methods are examined, providing a comprehensive grasp of the study's contributions to the academic community.

The researcher concludes the thesis in Chapter 6. The researcher respectfully recognises the constraints of the research, offering an open assessment of the difficulties faced and directions for further investigation. The chapter provides the reader with an overview of the overall significance of the research and helps them envision possible directions for future investigation. Each chapter is an attempt to develop the readers' understanding of lexical simplification in a comprehensive way and culminates with an investigation that both, advances the readers' grasp of the research area, and encourages future academics to explore this complex area in new ways.

Chapter 2

Literature Review

2.1 Lexical Simplification

Several studies on lexical simplification have helped advance this sub-field of text simplification over the years. Specia et al. (2012) conducted a shared task to devise automatic techniques for deriving simplifications of complex English words and phrases without compromising on their meaning. The evaluation metrics that were used were precision, recall, and F-score. The challenge involved using a dataset created by manually simplifying phrases from Wikipedia. A total of 17 systems took part in the shared task, each adopting a different strategy, including methods based on rules, machine learning, and hybrid techniques. The results demonstrated that the systems that combined machine learning-based and rule-based techniques performed the best. The subjective nature of simplicity and the necessity for context-awareness were two other issues that were covered in the paper's discussion of the challenges of lexical simplification. Siddharthan et al. (2015) proposed a supervised learning approach that used a combination of lexical and semantic features to identify difficult words in text and suggest simpler replacements. Similarly, Paetzold and Specia (2016a) proposed a neural network-based approach that used distributed representations of words to learn to simplify text. In addition to supervised learning, researchers have also explored unsupervised and semi-supervised methods for lexical simplification. For instance, Xu et al. (2016) proposed a method that used a combination of bilingual lexicons and word embeddings to automatically generate simplified versions of complex words. This approach was able to achieve comparable performance to supervised methods without requiring large amounts of annotated data. Another trend in lexical simplification research is the use of Neural Machine Translation (NMT) techniques. NMT models have been shown to be effective for simplification tasks because they can learn to translate complex sentences into simpler ones. For example, Wang et al. (2016) devised a framework for producing abbreviated versions of complex statements using a sequence-to-sequence NMT model. To train the NMT model, they used a parallel corpus made up of both complicated and simple texts. The training method for the NMT model included attention mechanisms and beam search decoding. Experimental analyses carried out on benchmark datasets revealed improvements in the readability and simplicity of the generated

simplified texts when compared to baseline methods, proving the efficacy of the suggested strategy.

There has been an increasing focus on the evaluation of lexical simplification models. Researchers have developed new benchmark datasets and evaluation metrics to assess the effectiveness of different approaches. For example, Vajjala et al. (2018) developed a benchmark dataset for lexical simplification and used it to evaluate a number of state-of-the-art models. The findings demonstrated that the models gave a low performance on the evaluation metric that focused on maintaining the meaning of the original sentence. Wubben et al. (2019) proposed a new evaluation metric that focused on the degree of simplification achieved by a model, rather than just the accuracy of word replacements.

Moreno et al. (2019) proposed a strategy for lexical simplification in order to support the accessibility guidelines of digital content. The authors argued that the use of jargon and sophisticated vocabulary in digital information could put impediments in the way of accessibility for those with disabilities and offer a three-step process for doing so. The first stage required recognising complex words and phrases, and the second involved utilising a thesaurus to look for simpler substitutes. In order to validate the suggested simplifications in context, a language model was utilised in the end. The authors tested their strategy using a dataset of web pages, and their results demonstrated that their technique could successfully simplify digital content while preserving its readability and significance. They claimed that their method could be used to make digital content more accessible, especially for those with disabilities. By recognising and substituting complex terms with simpler ones, Qiang et al. (2020) devised a framework that made use of language models that were pre-trained, such as BERT. The method produced cutting-edge results using benchmark datasets. The authors claimed that it can be easily adaptable to other languages. In order to simplify text for community-focused applications, Song et al. (2020) suggested a hybrid model that combined the benefits of rule-based and machine learning approaches. They generated initial candidates using rule-based simplification, and then utilised supervised learning to rank the choices according to context. More recently, Gooding et al. (2021) explored the subjective and context-dependent nature of word complexity, which makes it challenging to create entirely automatic systems for lexical simplification. The authors conducted a study in which participants were required to rate the complexity of words in various settings. They discovered that participants' perceptions of complexity varied greatly depending on the context and the reader. They argued

that community-driven approaches, in which the target audience participates in the simplification process, can better address this issue and that present automatic methods for lexical simplification do not take into consideration the subjective nature of word complexity. They also suggested that community-driven initiatives may be supplemented by computational techniques that could aid users in the process of simplification while also taking context and intended audience into consideration. Štajner (2021) provided an overview of numerous strategies for automatic text simplification, including methods based on machine learning, rules, and hybrid strategies. They also emphasised the need for a uniform evaluation framework as well as the need of evaluating and benchmarking various techniques. The study focused on the social benefits of text simplification in improving information accessibility for non-native speakers and those with cognitive and linguistic impairments. The author also studied the ethical ramifications of text simplification, with a focus on maintaining the author's intent and avoiding bias. The study emphasised both, the advancements in the field of automatic text simplification and the difficulties and moral issues that surround it. The author suggested that continued research and development in this area could help improve the accessibility of information for disadvantaged populations. Xu et al. (2021) developed a novel framework for lexical simplification based on neural machine translation with syntactic simplification. They used a parallel corpus of complex and simple texts to learn the translation patterns using a tree-based encoder-decoder model and a hierarchical attention mechanism. The proposed model outperformed several established models on multiple benchmark datasets. Glava and Tajner (2021) proposed a multi-task learning technique for lexical simplification. They trained a single model on several related tasks, such as word sense disambiguation, named entity recognition, and part-of-speech tagging. The proposed approach achieved good performance values on various benchmark datasets. Shardlow et al. (2022) created a new dataset of English texts that was labelled for lexical complexity. A variety of text genres, including fiction, academic writings, and news items, were included in the dataset named Complex 2.0. Annotators rated the difficulty of each word in the texts on a scale of 1 to 5. They also discussed the factors, such as frequency of the word, length of the word, and part-of-speech tags, that were used to estimate lexical complexity. The study compared the effectiveness of various machine learning methods and presented the findings of experiments using the Complex 2.0 dataset.

The results showed that the best-performing algorithm was a support vector machine, which achieved an accuracy of over 80% in predicting lexical complexity. The authors suggested that future work could explore the use of the dataset for other natural language processing tasks including assessment of readability, and text simplification. Gooding (2022) explored the moral

ramifications of text simplification, specifically the potential impact of the information communicated in a text. According to the author, text simplification can have unforeseen repercussions that may influence the accuracy and completeness of the information provided, even while it can be advantageous in boosting accessibility and comprehension for some audiences. The potential for prejudice or manipulation, the necessity for accountability and transparency, and the need of maintaining the text's original meaning and intent, were all mentioned as ethical issues that must be taken into account when language is simplified. The author also proposed a set of rules for moral text simplification that focused on the significance of including a variety of stakeholders and guaranteeing transparency. According to Gooding & Tragut (2022), current approaches to quantifying word complexity do not account for individual differences in reading ability and learning requirements. The authors proposed a novel way to gauge word complexity that took into account an individual reader's characteristics. In order to generate personalised models of word complexity, the authors introduced a framework for word complexity modelling that integrated linguistic word characteristics with data on individual reading skills. The models were created to give readers with various reading levels and learning requirements, personalised feedback and support. The findings showed that personalised models outperformed generic models that did not account for individual characteristics in terms of improving comprehension and readability. They argued that customised word complexity modelling could increase a reader's ability to read texts and understand them. The authors suggested that further research is required in order to improve the strategy and create tools and resources that can support personalized text simplification. Garcia (2022) suggested a technique for automatically making text simpler in order to adhere to requirements for cognitive accessibility. The suggested approach combined linguistic analysis and machine learning to identify and simplify complex words and phrases in text. The author identified and prioritised words and phrases for simplification based on a set of cognitive accessibility principles. The author reported the results of experiments using the proposed method to simplify a set of texts in different domains, including healthcare and education. The results showed that the method was effective in simplifying text while maintaining its meaning and coherence.

The author highlighted automated lexical simplification's potential as a tool for consistently supporting cognitive accessibility guidelines and suggested that this approach could be used to enhance the accessibility and comprehensibility of a variety of texts, particularly in fields where accurate and clear communication is crucial. Bhat et al. (2022) suggested a web browser extension that simplified text for readers with cognitive and learning difficulties. The extension allowed users to select different levels of simplification, from basic to advanced, and provided

visual cues to indicate the level of complexity of each. In order to offer a more complete solution for people with cognitive and learning problems, the authors recommended that their method could be expanded to additional languages and domains and combined with other assistive technologies. Alarcon et al. (2023) complied a corpus named the EASIER corpus. The corpus was created to simplify Spanish texts for individuals with reading difficulties, including those with intellectual disabilities and non-native speakers. It consists of 260 documents annotated by expert linguists, and was validated using Fleiss Kappa inter-annotator agreement coefficient and qualitative evaluation. Shardlow & Przybyła (2023) employed end-to-end simplification systems, unique language models, classification, and unsupervised and supervised approaches for lexical deletion. These techniques were tested on a new dataset named SimpleDelete, which was constructed using Simple English Wikipedia edit histories. The results show that unnecessary words can be deleted with unsupervised approaches. The study emphasised the importance of using deletion as a separate lexical simplification process.

Lexical simplification studies also discussed the challenges in the field. The challenge posed by words having multiple meanings is known as lexical ambiguity. Choosing the best replacement for a certain situation might be challenging. Research has been conducted by Agirre et al. (2016) on techniques for word disambiguation and replacement that align with the particular context in which a term is used. Lexical simplification model evaluation and training require annotated corpora. The challenge stems from the scarcity of extensive annotated datasets. To address this issue, the Complex Word Identification Shared Task (Yimam et al., 2017) and SemEval (Paetzold & Specia, 2016) worked to provide benchmark datasets and promote cooperation among researchers. Lexical simplification presents challenges when used on specialised or technical writing. In order to solve issues in particular domains, Narayan & Gardent (2014) investigated hybrid simplification strategies that make use of both machine translation and deep semantics. An ongoing challenge is adapting lexical simplification models to the intricacies of domain-specific languages and creating models for low resource languages. Beyond technological challenges, accessibility must also be taken into account. Hauser et al. (2022) highlighted the significance of multilingual text simplification and its practical applications in improving accessibility for linguistically diverse communities. To increase accessibility, Meli et al. (2024) created a lexical simplification method for Maltese text. Each step's performance was evaluated using an evaluation dataset, displaying encouraging outcomes. It was found that the method maintained the meaning of the sentence about half of the time while simplification. The researchers stated that even with positive results, more study

was required to solve unresolved issues. Qasmi et al. (2022) presented an unsupervised lexical simplification technique for complex Urdu text. They used morphological characteristics and word embeddings to produce simplifications from Urdu text. With manually constructed simplified corpora, the system achieved a BLEU score of approximately 80, and a SARI score of approximately 42 in automatic evaluations. They also discussed the outcomes of human assessments on simplicity, grammaticality, accuracy, and meaning preservation. Saggion et al. (2013) studied effects of various lexical resources and synonym selection techniques in a Spanish lexical simplification system. They made use of Spanish Open Thesaurus, Spanish EuroWordNet, and a combination of the two. They examined different approaches to word sense disambiguation, and devised a new evaluation framework that took into account the degree of ambiguity of the word. Based on word ambiguity levels, the researchers offered performance comparisons and suggested resources and disambiguation techniques. Ferrés et al. (2017) proposed a linguistic-statistical hybrid lexical simplifier that was multilingual, robust, and intended for use with Spanish, Portuguese, Catalan, and Galician. They investigated complex word detection using frequency thresholding over corpora frequency lists. Their work drew attention to problems with imbalanced corpora that impact frequency accuracy. They suggested ways to fix these issues, such as by manually altering frequency lists and fusing manually and corpus-based lists. They assessed the simplicity and appropriateness of the suggested based on judgements from native speakers. The study emphasised the need for advancements in morphological generation and word sense disambiguation algorithms. North et al. (2022) introduced ALEXSIS-PT, a new multi-candidate dataset for lexical simplification in Brazilian Portuguese. It contains 9,605 potential replacements for 387 complex words. Content from Brazilian newspapers was included in this dataset, which was assembled using the ALEXSIS technique for Spanish. By providing a broad benchmark dataset and a substantial quantity of ranked candidate substitutions, the researchers claim that ALEXSIS-PT closed major gaps in the lexical simplification liturature. Using this dataset, four models—mDistilBERT, mBERT, XLM-R, and BERTimbau—were evaluated for substitute generation. BERTimbau outperformed the others in terms of all evaluation metrics. Compared to the other models, which were trained on multilingual data containing several Portuguese dialects, BERTimbau's training on pt-BR data achieved better performance.

2.2 Readability Studies

Various readability studies have been conducted as part of which wordlists and readability formulae were invented.

2.2.1 Readability Studies using Word Lists

A commonly used word list that stated that shorter words are simpler to learn than longer ones was published by Thorndike in the Teacher's Word Book (Thorndike, 1921). Thorndike considered frequency to be a key criterion for evaluating difficulty. 10,000 English words are included in the list, which were organised by frequency and assigned to one of the grades between kindergarten and twelve. Researchers who studied the English language utilised this list to develop a readability metric (Lively & Pressey, 1923). The count of words and the number of distinct words in the text were also factors in the algorithm. It was noted that this list is made up of different spellings of the same word, indicating that the lemmas of the words were not taken into account when compiling the list. There were many proper nouns in the list as well. Proper nouns do not have any synonyms, hence including them in the list might not be useful when shortening words. Based on the frequency of the words, Edgar Dale developed a wordlist targeting children in grade 1 in 1941 using the International Kindergarten Union list and Thorndike's list (Dale, 1931).

2.2.2 Readability Studies using Readability Formulae

Experts have created readability metrics that indicate the complexity of a given text, or the level of formal education that a reader would require to possess for comprehending the information easily. For instance, a reading ease formula created by Flesch in 1948 (Flesch, 1948), the McCall-Crabbs Standard Test for English (McCall & Crabbs, 1925), a simplified Flesch reading ease formula that was developed by Farr et al. (1951), and the Gunning Fog index, which was based on the lessons from McCall-Crabbs (Gunning, 1952; McCallum & Peterson, 1982). These formulae have been used in various studies, for instance, in a study by Crossley et al. (2023). The researchers created the CommonLit Ease of Readability (CLEAR) corpus, which offers distinct readability ratings for around 5000 text excerpts along with details about the excerpt's genre, publishing year, and other metadata. Lahmar & Piras (2023) used

readability formulae to examine the readability of financial articles in a variety of finance publications and observed that scholarly articles on finance are difficult to read and comprehend. The Automated Readability Score was developed and evaluated in 1967 (Senter & Smith, 1967), and a significant association was established between the index and the text's grade level. The index indicated the general age range for readers who want to comprehend the content. Edward Fry created a graph-based technique in 1968 to determine the grade level of a given text (Fry, 1968). In the context of the US educational system, the graph focused on the English language. On a graph, the average amount of sentences and syllables were plotted for every hundred words. The grade was then determined based on the graph.

McLaughlin created a new readability formula the following year that he called SMOG (McLaughlin, 1969). McLaughlin stated that the number of polysyllabic words in a document determined how complex it was. The inclusion of words with three or more syllables in the formula was taken into consideration. Based on the reading scores determined by the reading test developed by Thorndike and McCall, the method generated a reading grade (Dransfield & McCall, 1925). The total number of years of formal training that the reader must have to comprehend the text was the output of this formula. A readability score of 1 meant that the text was difficult to understand, and a score of 12 meant it was easy to understand. In 1970, Coke and Rothkopf sought to determine the number of syllables in a word using mathematical evidence (Coke & Rothkopf, 1969). The Flesch-Kincaid Reading Ease formula was introduced in 1975 (Kincaid et al., 1975). The formula's output indicated the grade level in American schools that the reader must be in, to comprehend the material. The grade level of the text was also indicated by another formula developed by Coleman & Liau (1975) that same year. Several years later, the Strathclyde readability measure was proposed, which emphasised a word's frequency in 2006 (Weir & Ritchie, 2006). In 2007, Anula proposed an index to measure lexical complexity that used as its inputs, the lexical density and the index of terms whose frequency was low (Anula, 2007; Stajner & Saggion, 2013).

2.2.3 Readability Studies on Hindi

Bhagoliwal (1961) conducted one of the earliest studies on readability of Hindi text and evaluated the comprehension of Hindi short stories using English readability formulae. Hindi text was evaluated for readability using traditional English readability formula by Agnihotri & Khanna (1991), and the results demonstrated that surface features alone cannot predict a text's

readability. The readability of Hindi text was calculated using the model devised by Sinha et al. (2012) by utilising the average count of consonants, and the count of consonant conjuncts in the text. The count of consonant conjuncts and the average word length were both found by Sinha et al. (2014) to be accurate measures of text readability.

2.3 Lexical Parameters

Since the research does not focus on sentence or text readability, the researcher was unable to utilise the readability formulae mentioned in this chapter. However, the researcher aimed to establish a relationship between the word-level characteristics shared by these formulae and the word's complexity determined by a human intelligence task. The researcher considered features such as frequency, count of syllables, length, count of consonants, count of vowels, and count of consonant conjuncts, after separating the quantifiable attributes from the non-quantifiable ones. The researcher's goal was to ascertain the relationship between these characteristics and the comprehensibility of word in Hindi.

L.A. Sherman, an English professor, was one of the first researchers to conduct a study on readability in English. He observed that sentence length and word concreteness affect the reader's comprehension (Sherman, 1893). Sherman may not have provided a word list or formula, but there are many studies wherein various metrics were suggested to gauge the readability of a given book, most of which were written in the English language. In 1889, well-known Russian author Rubakin created a list of 1,500 words in his own language, claiming they were simple to understand. According to Rubakin, unfamiliar words and lengthy phrases make it harder for people to understand what is being said (Thorndike & Lorge, 1944). The different components that make up text's readability include its content, style, format, and organisation (Gray & Leary, 1935). However, the researcher's study primarily focuses on the lexical factors used in readability formulae.

The number of syllables, words, and sentences are all factors in the Flesch Reading Ease formula (Flesch, 1948). Researchers modified the coefficient values to adapt this formula to languages other than English. The Gunning Fog index (Gunning, 1952), the readability system developed by Edward Fry (Fry, 1968), SMOG (McLaughlin, 1969), and the Flesch-Kincaid Reading Ease formula (Kincaid et al., 1975) place a significant emphasis on the number of

syllables. Word length and frequency are additional variables frequently used in readability formulae (Senter & Smith, 1967; Coleman & Liau, 1975). The readability formula developed by Hazawawi et al. (2017) focused on the count of words in a phrase and the count of sentences in 300 words of a given text. In order to determine the readability of Vietnamese text, Luong et al. (2018) developed a formula that considered the average word length in terms of characters, the average length of sentences in characters, and the percentage of complex words that was ascertained using a pre-defined list of simple words. Length, consonant count, and consonant conjuncts were the factors taken into account when creating readability formulae for the Hindi language (Sinha et al., 2012; Sinha et al., 2014). Therefore, the goal of the research was to evaluate the significance of these lexical parameters in complex word classification models.

2.4 Dataset

This section contains details of lexical simplification datasets developed in several languages. The English Lexical Substitution Task of SemEval 2007 is one of the oldest studies on automatic lexical simplification datasets (McCarthy & Navigli, 2009). Native English speakers used a simpler word in place of the target word, in this task. However, this job was not intended to simulate the complex word recognition process. The researcher believes that by leaving out non-native speakers from the work, a bias may have been introduced. English lexical simplification was a SemEval 2012 task, and non-native speakers of English examined the corpus that was used in the study by McCarthy & Navigli (2009) (Specia et al., 2012). The annotators ranked the replacement terms for a group of target complex words in the dataset used for the 2007 SemEval task of English Lexical Substitution. The goal of this task was to come up with substitutes for the specified target term. The SemEval 2016 Complex Word Identification task concentrated on the challenge of identifying complex words (Paetzold & Specia, 2016b). Four hundred non-native English speakers participated in annotating the dataset. The task's goal was to annotate just one complex word in every sentence. The test set may have been biased because only one annotator annotated each sentence in the test set while 20 annotators annotated each of the sentences in the training set. The study focused on English, German, Spanish, and French and involved native annotators, as well as non-native annotators for the 2018 shared task on Complex Word Identification (Yimam et al., 2017b). The CWIG3G2 dataset used for this work was generated by Yimam et al. (2017a). The annotators assigned the likelihood that the supplied target word was complex, and they modelled complex word identification in the form of a binary classification problem and a probabilistic

classification task. 47 non-native annotators and 134 native annotators participated in the task. The organisers instructed the annotators to focus on assumed complexity by assuming that the intended readers were either children, people with language disorders, or language learners. Maddela and Xu (2018) constructed a dataset for a neural ranking algorithm for measuring readability. They used the Google 1T Ngram Corpus and selected the most popular 15,000 English words (Brants, 2006). Eleven non-native English readers annotated the words on the Likert scale (Likert, 1932). Only correlated terms were used in the annotations, and the Pearson's correlation coefficient was obtained between the annotation made by each annotator and the average of the annotations provided by the other annotators. The average of the ratings each word received, served as the final evaluation.

CompLex 1.0 and Complex 2.0 are two datasets produced by Shardlow et al. (2022). The sentences in the dataset in CompLex 1.0 were annotated according to their level of complexity. However CompLex 2.0 was an enhancement, because the count of instances and the count of annotations for each instance was more in CompLex 2.0 as compared to CompLex 1.0. The English Bible text, Europarl (European parliament proceedings), and biomedical publications make up this dataset. Each word was annotated using a 5-point Likert scale that represented complexity. Similar to the procedure used by Maddela & Xu (2018), they determined the correlation between the annotation given to a word by an annotator and the average number of annotations it received, to ensure the quality of the annotations. They also determined the relationship between the frequency of a word and the annotations it received. The complexity value of a word was determined by averaging its annotations on a normalised scale. They assumed a relationship between each participant’s annotation and the average of all the other annotators' annotations. They assigned a value for complexity that was averaged over the annotations. It is unclear if the complexity value offered by an annotator was determined by comparison with the familiarity of other unrelated terms or by the familiarity of other related, or synonymous words. LSeval (De Belder & Moens, 2012), the English CW Corpus (Shardlow, 2013), and other datasets with complex terms and their simpler equivalents are also available. SIMPLEX-PB targeting Portuguese language (Hartmann et al., 2018), BenchLS (Paetzold & Specia, 2016c), and SNOW E4 (Kajiwara & Yamamoto, 2015) are examples of tools for languages other than English. Kai et al. (2023) present two multilingual datasets, ALEXSIS++ and ALEXSIS+. While ALEXSIS++ is an English-only dataset, ALEXSIS+ covers English, Portuguese, and Spanish. These datasets, which were obtained from news corpora, have over fifty thousand sentences annotated with cosine similarity to the original complex words and

sentences. The researchers used the dataset to generate candidate substitutions for the specified languages. Valentini et al. (2023) used samples from the stories generated by large language models in their experiment to create a collection of lexical simplification instances targeted at children. In order to investigate Japanese lexical complexity, Ide et al. (2023) created the first Japanese lexical complexity prediction dataset. The dataset offers distinct difficulty scores for Chinese and Korean native readers. The study showed the efficacy of a BERT-based solution for Japanese lexical complexity prediction in the baseline experiment.

In several of the tasks requiring annotation, the complex word was predetermined. Ranking the less complex alternatives to the complex word was anticipated of the annotator. In other scenarios, the participants were either native speakers or they were given instructions to assume specific things about the intended audience when they annotated. The lack of a dataset for Hindi lexical simplification and complex word identification in Hindi text, as well as the difficulty in finding a dataset with lexical properties of words, inspired us to design an annotation study and construct a dataset. The requirement for a new dataset is also explained by the linguistic variations between Hindi and other languages for which datasets exist. Additionally, since there has never been a study that links a word's frequency in Hindi texts to its complexity, the researcher could not infer that frequency is an important factor.

The goal of the researcher’s study included conducting an annotation task, analysing the annotations, and creating a dataset, in order to detect complex words in a particular Hindi text.

2.5 Stop Word Identification

Stop words are words in a sentence that exist only for grammatical purposes and do not add anything to the knowledge you can obtain from the text (Joshi et al., 2012). Therefore, if these stop words are found and eliminated before utilising the text for a task, the task's performance may be enhanced. The significance of including the elimination of stop words in the pre-processing stage for text processing tasks has been the area of interest of numerous studies (Jha et al., 2016; Silva & Ribeiro, 2003; Na & Xu, 2015).

Rule-based techniques were employed in most of the early studies on stop word identification. Lists of frequently occurring stop words were compiled based on linguistic awareness or

frequency data (Salton & McGill, 1983). These methods were effective, but they were limited in the languages and subjects they could be used in. The statistical distribution of words was taken into consideration when frequency-based techniques gained popularity. Words that appeared frequently in a corpus were considered potential stop words. In order to detect stop words, Manning et al. (2008) used statistical metrics such as term frequency-inverse document frequency, also known as TF-IDF. The TF-IDF measure is a quantitative measure used to evaluate the importance of a word in a document with respect to a corpus, that is, a collection of documents. Achsan et al. (2023) also used TF-IDF to identify stop words. Their work addressed a major challenge in stop word extraction for the Indonesian language as it did not require any prior linguistic understanding. They were able to extract all stop words found in all documents and used an automatic cut-off technique to choose the top-ranked stop word candidates.

In order to train classifiers to classify between stop words and content words, supervised models frequently used labelled stop word datasets (Yang & Pedersen, 1997). Kucukyilmaz & Akin (2023) devised an automatic feature-based supervised machine learning method for stop word detection. They conducted experiments to verify the viability of the suggested method and compared the outcomes with a list of common English stop words. They tested the method with formal as well as informal text. The findings indicated that the suggested method was capable of treating the dialectal changes and produced encouraging results. However, according to Baker & McCallum (1998), stop words should be identified using unsupervised techniques such as clustering algorithms, which were based on contextual patterns. In light of the significance of domain-specific stop words, researchers looked into methods for customising lists of stop words to certain domains. Makrehchi & Kamel (2017) presented a novel approach that made use of a huge labelled corpus to generate stop words that were specific to a certain domain, as opposed to traditional methods that depended on high or low document frequencies or lists of typical stop words. They introduced an algorithm for creating lists of stop words that are both domain-specific and general. The study defied the notion that efficient term ranking metrics also excel at choosing stop words. Through a comparative study, the authors showed that their innovative approach gave more promising results, ensuring minimal information loss by effectively filtering out a significant portion of stop words.

In order to create a stop word list in the Hindi language containing a total of 275 words, Larkey, Connell, & Abduljaleel (2003) extracted stop words manually using parts of speech like

conjunctions, prepositions, pronouns, and other from two news-based corpora. Rani & Lobiyal (2018) presented a technique for automatically generating lists of stop words that matched with the twenty most popular stop word lists extracted from four publicly accessible lists. Although the researcher discovered studies by Lo, He, & Ounis (2005), Zou, Wang, Deng, Han & Wang (2006), Hao & Hao (2008), Yao & Ze-wen (2011) and Alajmi et al. (2012), and other studies that made their stop word lists available for public consumption that were centred on automatic stop word generation (Kaur & Saini, 2016) , the researcher was unable to find a publicly available list of stop words in Hindi that was based on multiple corpora.

The inconsistent use of words in the lists, is another issue with many lists based on either one or two corpora (Murphy, 2012). Kaur & Saini (2016) compiled a stop word list with 256 words from writings and poems in Punjabi. Lemmatizing the words reduced this list to 184 distinct words. The researchers performed word lemmatisation after determining which words were stop words rather than lemmatizing all the words first. Rakholia & Saini (2016) manually compiled a list of Gujarati news corpus stop words. They also looked for patterns among the stop words with the same part-of-speech tags, but they were unable to uncover any. By assessing the word frequencies in the text, another study developed a technique that uses threshold to recognise stop words in Sanskrit text (Raulji & Saini, 2017). Nouns were excluded from the list by the researchers on the grounds that they could not be used as stop words. Sagar & Saini (2023) used three different approaches to produce a list of 279 Koshur stop words. 110 stop words were taken out of Koshur stories, poems, publications, and internet resources. After eliminating redundant and irrelevant words, they translated and combined stop words from Punjabi, English, and Hindi to create a list of 147 stop words. They added 22 new stop words and achieved 89% accuracy in their frequency counts of 27 Koshur poetry and 9 folktales. Three language specialists verified the final list, confirming 100% accuracy and requiring no modifications. Through this research, the researcher aimed to examine the veracity of this supposition. The researcher also observed that a majority of the research the researcher studied used news stories as their data source.

The majority of these studies focused more on creating list of stop words rather than stop lemmas. This would lead to the construction of a list that was made up of morphological variations of words rather than the root word itself, and this in turn, could affect the quality of the task negatively, because the list was not robust and defeated the purpose for which it was

formed. Depending on the language's morphological complexity, the inclusion of many word forms of the same word in the stop word lists raised additional concerns.

After studying the existing works, the researcher framed the objectives of the stop-word identification study as follows:
a) To ascertain the interchangeability of the current stop word lists made available to the general public.
b) To compile a comprehensive collection of stop lemmas regardless of the intended domain

2.6 Complex Word Identification

The earliest research on complex word identification may be found in the field of medicine, where researchers used term occurrence to predict how familiar a medical term would be, and showed that personalised appraisal is achievable provided the models take the reader demographics into consideration (Zeng et al., 2005). Another study examined the use of WordNet sense counts, MRC Psycholinguistic Database familiarity characteristics, and corpus frequencies (Brown, 2005; Wilson, 1988; Elhadad, 2006).

The researcher learned from the literature that there were three ways to identify complex words. The initial strategy was to simplify every word in the text by assuming that they were all complex words (Devlin, 1998; Bott et al., 2012). The disadvantage of this approach was that it took time, complicated previously simple statements, or completely changed the statement's meaning. Another approach to evaluating a word's complexity relied on a threshold that was frequently dependent on word frequency (Hsueh-Chao & Nation, 2000; Elhadad, 2006; Biran et al., 2011). A third strategy framed the task of identifying complex words in the form of a machine learning challenge (Zeng et al., 2005; Biran et al., 2011). In early literature, the complexity of a word was determined by readability criteria based on the familiarity of a word (Dale & Chall, 1948), or syllable count (Gunning, 1952; Mc Laughlin, 1969). The CW corpus, which Matthew Shardlow constructed in 2013, contained 731 instances that were taken from Simple English Wikipedia updates (Shardlow, 2013). A classifier that categorised English words could be trained using this dataset.

A significant amount of work undertaken on this topic was submitted as entries to tasks organised under the SemEval series of workshops. McCarthy & Navigli (2009) organised the first such collaborative task. The goal of the task was to identify a context-appropriate replacement for a given word. Subsequently, a 2012 task was set with the goal of ranking the context-relevant word replacements in order of increasing complexity (Specia et al., 2012). Four participants annotated the trial dataset, whereas five participants annotated the test dataset. The participants were given synonyms with comparable levels of complexity. They had to select a minimum of three complex terms on each screen. The goal of ranking these terms according to complexity was then presented. The participants were asked to rank the words by presuming that the text would be read by people who were not native English speakers. These exercises were completed in varied order for various participants. The researcher found that the time needed to rank the words was less than the time needed to designate a word as complex. This might be because the task requiring comparison judgment did not require the participant to read a text. Since the participants' backgrounds were taken into account when recruiting them for the study, it was also discovered that the agreement values were higher than in earlier studies. The inter-annotator agreement was determined using pairwise rank comparisons and the Kappa index.

Paetzold and Specia organised yet another task in 2016 (Paetzold & Specia, 2016b). 9200 sentences were annotated by 400 non-native English speakers. The organisers chose the target words for this task. The participants were asked to annotate any word that they were unable to understand even if they were able to comprehend the meaning of the sentence in which the word appeared. Twenty non-native participants annotated the training set, which was composed of 200 sentences. The remaining 9000 sentences contained sets of sentences, wherein one participant annotated each set. Ten-sentence sets of each type were created. Twenty people annotated one sentence each. The agreement value was low because of the diversity in the participants' backgrounds. A word was marked as complex even if only one participant designated the word as complex. 3,854 unique words out of 35,958 words were annotated as complex, indicating that the dataset was skewed. Based on the solutions submitted for the task, the researcher found that systems using decision trees and random forests produced good results, and that ensemble methods outperformed other machine learning methods, particularly neural approaches. The frequency of a word stood out as the most important characteristic to distinguish a complex word.

In a similar task conducted in 2018, participants were asked to mark a term as complex if they believed it to be difficult for young readers, non-native speakers, or individuals with language disorders to comprehend (Yimam et al., 2018). They used the CWIG3G2 dataset (Yimam et al., 2017a), and the work included French, German, and Spanish in addition to English. Additionally, multi-word expressions were annotated. The count of native and non-native participants was not consistent between languages, and the proportion of native and non-native speakers annotating for a single language was unbalanced. Complex terms were not marked before the annotation process, in contrast to the preceding task. A word was considered complex if at least one participant assigned a complex rating to the word. The only features considered by the baseline system were the frequency and word length. Macro-averaged F1 and accuracy were utilised to assess the models. Since the participants comprised both, native as well as non-native English speakers, the percentage of participants who agreed that a word was complex in English, was low. When compared to non-native speakers, it was found that native speakers agreed more frequently. German was shown to have a greater agreement percentage among non-native speakers than among native speakers. Despite the fact that the majority of the annotators were native Spanish speakers, the Spanish annotation task produced lower agreement values when compared to the results of the English and German annotation tasks. It was also found that classical machine learning techniques outperformed neural network-based techniques. Figure 2.1 and Figure 2.2 provide an overview of the characteristics employed in the CWI Shared Task 2016 and CWI Shared Task 2018, respectively.

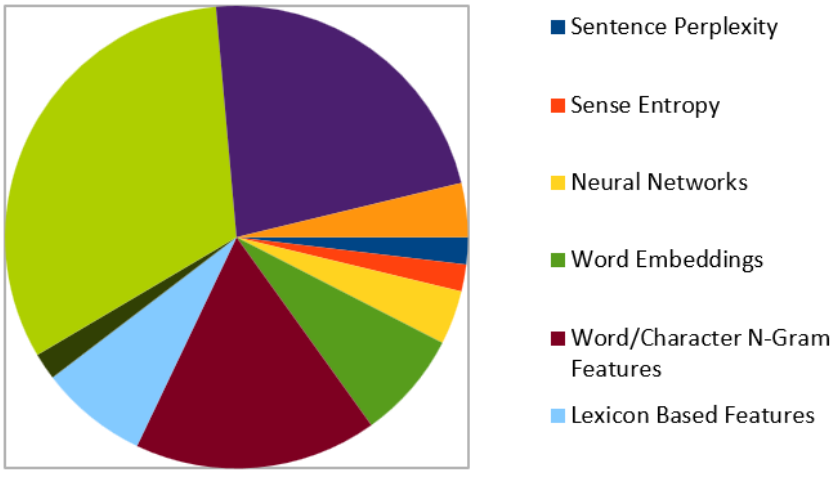

Figure 2.1: Summary of features employed in Complex Word Identification SemEval-2016 Task 11

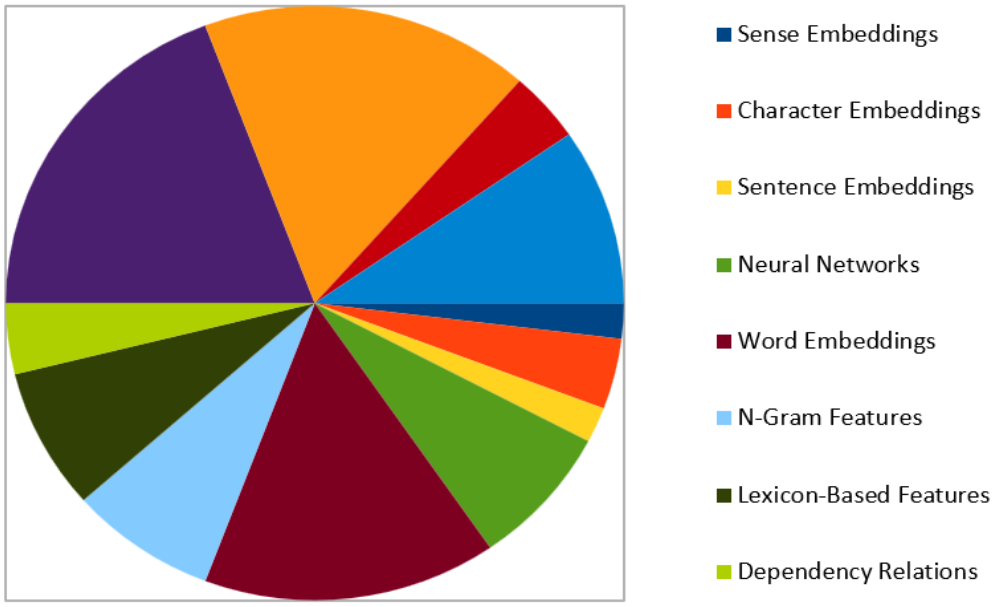

Figure 2.2: Summary of features employed in Complex Word Identification Shared Task 2018

As can be observed, submissions to both challenges have a strong emphasis on lexical and semantic aspects as well as word frequency. Another study found that compared to terms that were tagged as simple, complex words had a greater average word length (Zampieri et al., 2017).

Ortiz-Zambrano et al. (2023) extracted 23 linguistic features and combined them with encodings generated by different language models. They found that the performance of transformer-based models can be improved by integrating linguistic information automatically derived from texts. North & Zampieri (2023) discovered that non-native speakers' perception of lexical complexity was more influenced by elements such word frequency, length, syllable count, prevalence, familiarity, and the number of derivations than native speakers. The results were explained by a number of ideas in applied linguistics, and a binary classifier was created to differentiate between spelling mistakes made by native and non-native English speakers. Alghamdi et al. (2023) conducted a study to determine whether current lexical complexity metrics could be expanded to predict how difficult videos would be perceived by second language learners. Regression models were created for explaining and predicting difficulty using lexical complexity, utilising a corpus of 320 instructional films. The findings validated important lexical complexity features in assessments of video difficulty. The highest variance in the evaluation of video complexity was explained by lexical frequency indices.

2.7 Use of Word Embeddings in Complex Word Identification

Kuru (2016) worked on identifying complex words in English text. They tested with random embeddings, several forms of pretrained embeddings, and a word substring as a feature. They employed a linear support vector machine classifier that was trained using grid search hyperparameter tuning and five-fold cross validation. The measurement metric for evaluation was G-score. The model trained on random embeddings underperformed the pretrained embeddings. Then, they combined two distinct embedding types and used them as features of the target word. These were combined with the target word's suffix, prefix, and character n-grams as features. 5-fold cross validation and random search were used to train and fine-tune the model, respectively. They found no significant improvement when they compared the target word's context-sensitive form to its context-agnostic representation. The system performed well when it was trained using a concatenation of several embeddings and substring features. Sheang (2019) developed a technique in order to find complex words and phrases in texts written in English, Spanish, and German. They combined GloVe embeddings with other features such as frequency, count of syllables, length, count of vowels, part of speech tag, dependency, stop word, and tf-idf value (Pennington et al., 2014). All feature values were adjusted to fall within the range of 0 and 1. Every feature was transformed into a matrix that represented the target word and other words in its context. They used Convolutional Neural Networks (CNN) with max pooling and ReLu activation to train the model. Three fully connected layers received the output with dropout. The last fully-connected layer employed softmax activation, which produced an output of 1 or 0, denoting complex and simple, respectively, depending on the input. They used stochastic gradient descent using the Adam optimizer for training. The evaluation metric was the macro-F1 score, and the loss function was weighted cross-entropy. They implemented grid search hyperparameter tuning to adjust the model. Aroyehun et al. (2018) compared feature-engineered tree models and a convolutional neural network model for recognising complex words and sentences. English and Spanish were the languages they focused on. The English dataset showed good results for both of these strategies. They employed a variety of linguistic features, including fasttext embeddings for Spanish and word2vec pretrained embeddings for English (Mikolov et al., 2013). (Bojanowski et al., 2016). The similarity between the target word or phrase and its context was also used as a characteristic. They achieved satisfactory performance with tree-based ensemble models,. They normalised the vectors using CNN so that their values were within the range of 0 and 1. The context of the target word and its embeddings were both represented. They combined fully-

connected layers with a ReLu activation function and max pooling. A linear activation function was utilised in the first two fully-connected layers, whereas a sigmoid activation function was used in the third layer. This produced a result between 0 and 1. They established a threshold of 0.5 and used it to determine whether the result was simple or complex. When training the network, they utilised RMSprop as the optimizer and binary cross-entropy as the objective function. De Hertog & Tack (2018) made an effort to find complex words in texts written in English and Spanish. They made use of word length, frequency, word2vec word embeddings, character embeddings, and psycholinguistic characteristics. By joining the word embeddings within a sentence, they were able to extract topical information. These were fed into the three-layer, fully connected network. They employed mean squared error rate for the probability task and binary cross entropy as the loss function for binary prediction. They discovered that character embeddings performed better in predicting word complexity as compared to word embeddings. They observed, in line with earlier studies, that neither context nor topic information contribute significantly. Word2vec embeddings, POS tags, frequency, ambiguity counts, word length and syllable counts, and the similarity value of the target word with the words in context, were employed by Sanjay & Soman (2016) as input to an SVM classifier.

An SVM classifier that determines the level of difficulty of words in Japanese text was trained in a different study. They employed word frequency, character frequency, word2vec embeddings, part of speech tag, and word frequency as features (Nishihara & Kajiwara, 2020). Word embeddings were proven to be a useful feature for predicting the complexity of French words (Soler et al., 2018). Using shallow information including word length and position in the sentence, dependency tree features, semantic features, and word2vec representations of the phrase, AbuRa'ed and Saggion (2018) constructed a random forest classifier. Yang et al. (2023) devised a novel supervised method to solve the lexical complexity prediction problem as a single-label multiclassification problem by utilising word embeddings. Three datasets in the languages of English, Traditional Chinese, and Japanese were used to assess the prediction models. The outcomes demonstrated that for the English dataset, SVM with fastText embeddings can reach the maximum accuracy of 66.23%. The Traditional Chinese dataset yielded the maximum accuracy of 53.84% when using SVM with GloVe embeddings. For the Japanese dataset, SVM with Word2Vec embeddings achieved the best accuracy of 49.96%.

As word embeddings were proven to be useful for identifying complex words, the researcher decided to test context-neutral pretrained embeddings based on the findings from the literature.

An objective of the research study was to ascertain whether a word in a Hindi sentence is simple or complex, regardless of the context in which a it appears. The researcher attempted to reproduce the system that produced the best results in shared tasks. The researcher decided against replicating the systems due to the absence of comparable resources in Hindi, such as psycholinguistic databases, simple word lists, and the Simple Hindi Wikipedia, i.e., resources that were used by previous studies. The study's main contribution is the methodology the researcher developed, which normalises features so that the system does not compare unrelated words to gauge complexity, which was ignored by earlier work in other languages.

2.8 Word Sense Disambiguation

One of the most crucial tasks in natural language processing is word sense disambiguation. The task of word sense disambiguation refers to determining the correct meaning of a word in a given context. The field has advanced drastically, and numerous studies have helped create efficient word sense disambiguation systems. Researchers across the world have developed diverse strategies to address the complex issue of word sense disambiguation. These methods can be divided into a number of groups, including two that are supervised and knowledge-based methods (Borah et al., 2014). Knowledge-based approaches use tools such as a lexicon or a sense-inventory, whereas supervised approaches use a manually annotated dataset. Supervised word sense disambiguation has frequently demonstrated superior performance as compared to knowledge-based methods (Raganato et al., 2017).

Earlier word sense disambiguation methods used annotated corpora for training and supervised learning. One of the foundational studies is Yarowsky's unsupervised learning approach (Yarowsky, 1995). Techniques such as SemCor and Senseval, which rely on sense-annotated corpora, became standards for creating reliable word sense disambiguation systems (Gale et al., 1992; Miller et al., 1993). WordNet and other external lexical resources are utilised by knowledge-based approaches. The method developed by Lesk (1986) is a noteworthy example. It compares word sense overlaps in dictionary definitions. Further research investigated knowledge-rich features and graph-based algorithms, showing the value of leveraging semantic linkages from lexical resources (Navigli, 2009). Researchers studied unsupervised methods to solve the lack of annotated data. Agirre et al. (2007) and Keller & Lapata (2004) explored clustering algorithms, sense induction techniques, and distributional semantics models. The

goal of these methods was to reduce the dependence on labelled data so that word sense disambiguation may be used in languages and contexts considering a scarcity of annotated resources. Studies on sensory embeddings and neural architectures for word sense disambiguation increased with the development of neural networks. Semantic similarities between words and senses were obtained by embedding techniques such as word2vec and GloVe (Mikolov et al., 2013; Pennington et al., 2014). By learning contextual representations for words in context, neural network topologies, such as transformers and recurrent neural networks, produced state-of-the-art results (Huang et al., 2012; Devlin et al., 2018). By modifying word sense disambiguation models for various linguistic situations, studies examined the transferability of knowledge from resource-rich languages to resource-poor languages (Gonzalez-Agirre et al., 2012). Research in this area has advanced significantly owing to manually annotated corpora (Pasini, 2021). With the introduction of pre-trained language models, the discipline advanced even further (Huang et al., 2019; Loureiro & Jorge, 2019; Vial et al., 2019; Bevilacqua & Navigli, 2020; Blevins & Zettlemoyer, 2020). The two forms of word sense disambiguation tasks that have been experimented with, by researchers are the all-words task and the lexical sample task. The all-words approach seeks to identify all the content words in a given text, whereas the lexical sample task consists of annotating the correct sense for a group of terms used in various circumstances (Iacobacci et al., 2016).

Saeed et al. (2019a) developed a corpus aimed at the Urdu lexical sample word sense disambiguation task. There were 7,185 sentences in this corpus, along with 50 annotated words containing 9 verbs, 30 nouns, and 11 adjectives. With the aid of a dictionary that was utilised to recover the senses, three annotators annotated these terms. Here, the third annotator was employed to settle the conflicts between the other two annotators. They used recall, precision, accuracy, and F1-score, to assess the model's performance, achieving a respectable weighted-Kappa score of 0.82. Wiedemann et al. (2019) were successful in achieving satisfactory results with contextualised embeddings for word sense disambiguation. Rouhizadeh et al. (2021) published a sense-annotated corpus in the Persian language by utilising an all-words method for word sense disambiguation. 3,371 annotated words including verbs, adverbs, nouns, and adjectives were present in the corpus. Using an all-words technique, Saeed et al. (2019b) produced a sense-annotated corpus with 466 categories and 856 manually tagged tokens. Three native Urdu speakers collaborated on the annotation; the third speaker assisted in resolving disagreements between the other two. They selected a document with 2,306 content words that, after pre-processing, was reduced to 466 ambiguous terms. 5,042 words and 252 phrases

composed the final corpus. They evaluated the corpus's quality using accuracy. With the aid of manual annotation and three supervised classifiers, decision tree, k-nearest neighbours, and naive Bayes, Walia et al. (2018a) presented 100 sense-tagged words in Punjabi (Murphy, 2006; Fix & Hodges, 1989). Using a dictionary and manual annotation, Pal et al. (2019) separated words in Bengali text, using the percentage of accuracy as the performance parameter. To distinguish between different Punjabi words, Walia et al. (2018b) employed a sense-tagged corpus using K-Nearest Neighbors. Using the Hindi WordNet (Bhattacharyya, 2008) and a genetic algorithm that exclusively disambiguated nouns, Athaiya et al. (2018) examined word sense disambiguation in Hindi text. To distinguish between Hindi words, Kulkarni and Rodd (2022) employed different iterations of the Lesk algorithm (Gautam & Sharma, 2016; Tripathi et al., 2020). By assessing the sense-gloss overlap, Karuppaiah & Vincent (2021) developed an unsupervised approach for disambiguating Tamil words' senses. Hadiwinoto et al. (2019) investigated the use of contextualised pre-trained embeddings for word sense disambiguation. They claimed that their approach outperformed approaches based on neural networks, feature-based analysis, and context-agnostic embeddings. Mishra & Jain (2023) proposed a method that used the long short-term memory neural network along with word2vec embeddings to perform word sense disambiguation in Hindi text. The method found the correct sense according to the context of the target word in the Hindi sentence. Their model produced an F1-score of around 62%. Jha et al. (2023) proposed the use of a weighted graph with nodes indicating the meanings of words that occur in the context of ambiguous phrases, and edges indicating the relationships between them. They employed a random walk-style approach to determine which sense of a polysemous word was more suited in a particular context and leveraged semantic similarity computed from the Hindi WordNet (Bhattacharyya, 2008) to provide weight to edges. Twenty polysemous nouns from a sense-annotated dataset were used for the evaluation. Their overall accuracy of 63.39% was higher than previously published research on the same dataset. Kaddoura & Nassar (2023) devised a method to determine the overlap between dictionary definitions and contextual information by using similarity measurements. Their technique used a pre-trained BERT language model to provide efficient Arabic word disambiguation. New characteristics were added during training to improve the model's capacity to differentiate between various word senses. Combining several BERT models resulted in an ensemble model architecture that enhanced classification performance. Scarlini et al. (2020a) utilised BERT to build the embeddings of word senses across several languages. However, Hindi was not one of them; they focused on the nominal senses of five distinct languages. By training embeddings on only English text, they later expanded on this work and used BERT and mBERT models to

create a semi-supervised technique for disambiguating senses in other languages (Kenton & Toutanova, 2019). (Scarlini et al., 2020b). They published a dataset comprising multilingual ARES word embeddings and their BabelNet-retrieved synset IDs (Navigli & Ponzetto, 2012). In addition to outperforming other cutting-edge word sense disambiguation systems, this system cleared the way for multilingual word sense disambiguation in languages for which the senses are available in BabelNet. Even though Hindi was not included in the study, the researcher was determined to test the use of embeddings to determine how well they can distinguish between words in Hindi text. With an intent to determine the correct sense of a target word in a Hindi sentence, the study used ARES-multilingual embeddings for fetching sense embeddings together with a variety of pretrained embeddings. The researcher's goal was to identify the most effective language model and evaluate its performance using a manually annotated dataset, to accomplish Hindi word sense disambiguation using ARES-multilingual embeddings.

Chapter 3

Methodology

The methodology is divided into five sections – corpus and stop word list creation, dataset creation, study of lexical parameters of classical readability formulae, model creation, validation and selection, word sense disambiguation and synonym selection.

3.1 Corpus Creation

This section highlights the construction of a corpus containing text that falls under the aesthetics domain, and provides insights into the metadata of the collated corpora. The researcher set out to create a corpus in order to create a dataset that would be presented to annotators for creating a labelled dataset. This corpus would also be used, in combination with other corpora, to calculate the frequency of a given word.

The researcher utilized web scraping techniques to gather biographies, stories and novels from different sources, including a website that is dedicated to the well-known Hindi novelist Premchand's literary work[1], an online library run by the Digital Library of the Bhandarkar Oriental Research Institute[2], and the online magazine of the Mahatma Gandhi Antarrashtriya Hindi Vishwavidyalaya, Wardha[3]. The researcher used Scrapy, which is an open source software used to extract content from these websites. Additionally, the researcher extracted textual content from PDF files of literary work that was not present in the specified portals. However, encoding errors prevented the utilization of this extracted text. The researcher faced numerous obstacles when navigating the process of creating the aesthetics corpus, which were made worse by the limitations of having restricted access to publicly available content.

[1] https://premchand.co.in

[2] https://borilib.com

[3] https://hindisamay.com/

Because of the inherent complexity involved in creating this corpus, it was not possible to compare its size to articles that were sourced from other repositories, such as Wikimedia database dumps[1]. In spite of these obstacles, the aesthetics corpus proves to be a unique and invaluable asset. When combined with other corpora, its inherent worth is amplified and offers a wealth of information for upcoming research projects.

The preprocessing approach included basic tokenization, sentence splitting operations and data cleaning. The researcher removed special characters, Latin numerals, and English tokens. Here, the researcher did not segment joined words.

A crucial step in the research process was the extraction of lemmas, which are the basic root forms of words. With the help of the stanfordnlp Python package (Zeman et al., 2018), this task was completed accurately and effectively. For numerous natural language processing (NLP) tasks, such as sentiment analysis, tokenization, natural language generation, part-of-speech tagging, and more, the stanfordnlp library provides pre-trained models and tools, which greatly contribute in natural language processing studies.

Table 3.1 provides details about the word and lemma count in the corpus, and the corpora that the researcher obtained from various sources. The researcher generated the lemma by extracting the root form of the word using the stanfordnlp Python package (Zeman et al., 2018). In the table, "LR" refers to the language resource the researcher created and the resources that were used in the study, where "LR-1" refers to the corpus the researcher created. The specifics of these language resources, along with a detailed enumeration, can be found in the documentation provided in Appendix I. The richness of the corpora employed in this study is evident in their coverage, spanning ten distinct domains that collectively span a period of more than a century. This adds a historical depth to the exploration, shedding light on the evolution of language across diverse contexts over time.

[1] https://dumps.wikimedia.org/enwiki/

Table 3.1: Corpora metadata

S.No.	Source	Count of Unique Words	Count of Unique Lemmas	Domain
1	LR-1	145,508	118,266	Aesthetics
2	LR-2	21,335	17,159	Entertainment
3	LR-3	119,313	102,201	Not available
4	LR-4	2,330	1,851	Entertainment, Education, Judicial, Aesthetics, and Administration
5	LR-5	21,826	18,220	Tourism
6	LR-6	39,351	32,074	Entertainment and Agriculture
7	LR-7	35,018	28,645	Entertainment, Agriculture, Sports, Politics, Aesthetics, Literature, Economy, and Religion
8	LR-8	20,430	16,673	Health

The corpus compilation effort culminated in the assimilation of a substantial volume. The researcher was able to collect 978 articles from the predetermined sources. This collection comprises a diverse array of literary genres, including biographies, short stories, novels, and non-fictional material, amplifying the breadth and depth of this linguistic exploration.

However, in the pursuit of metadata exploration and documentation, a subset of 164 items within this compilation of 978 items, presented a challenge. Despite exhaustive efforts, the metadata for these specific items presented ambiguities associated with historical and diverse data sources. This observation of data limitations serves not only to transparently communicate the extent of available information and the scope of the research, but also sets the stage for future research considerations and potential refinements.

Fig. 3.1 depicts the distribution by states of authors whose works were incorporated into the corpus. As shown in the figure, most of the work is linked to authors from Uttar Pradesh, which

is a state in northern India. The metadata also revealed that just 4.84% of the included publications were written by women. For this corpus, this count was 0 before independence. Although it cannot be compared to the number of male authors, the number did rise as the years went by.

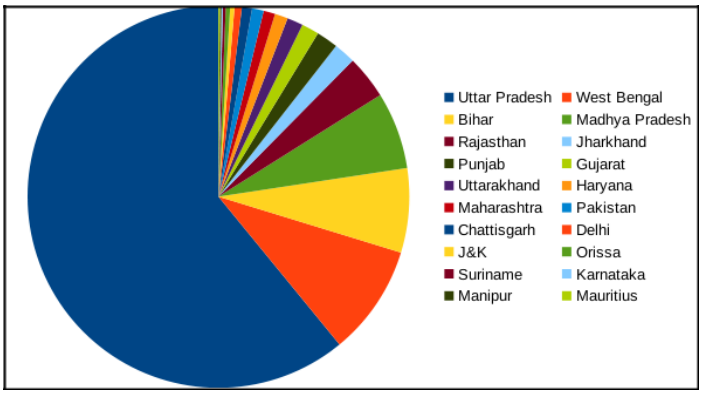

Figure 3.1: Distribution of authors by state

Although the researcher was unable to locate any information about less accomplished authors, the researcher assumed that they are not amateur authors, as the institution decided to showcase their work on its website.

3.2 Stop Lemma List Creation

The researcher attempted to answer the research questions that were formed after reviewing the literature.

RQ1: Are the top 10 stop words in all lists of stop words available the same, regardless of the position of the words in the lists?

The null hypothesis is as follows:

H_0: Regardless of the order in which the words are present in the lists, the top 10 stop words for all stop word lists are the same.

H_A: Regardless of the order in which the words are present in the lists, the top 10 stop words for all stop word lists are distinct.

The researcher created lists of top 10 stop words from multiple resources, as can be seen in Table 3.2.

Table 3.2: Top 10 stop words present in each source

Language Resource	Stop Words									
1	है	के	में	की	और	से	का	को	नहीं	तो
2	है	के	मैं	नहीं	में	हैं	एक	आप	और	लिए
3	के	है	में	की	से	और	का	को	हैं	पर
4	के	में	की	और	लिए	हैं	है	से	का	को
5	के	है	में	और	की	से	एक	का	हैं	को
6	के	है	में	की	का	से	और	को	हैं	भी
7	के	है	में	की	का	से	को	और	हैं	ने
8	है	के	में	से	की	को	का	हैं	और	हो
9	में	है	हैं	नहीं	लिए	गया	तथा	अपने	कुछ	साथ
10	जैसा	मैं	उसके	कि	वह	था	के लिए	पर	हैं	साथ
11	की	की	एक	तक	में	है	आप	कि	यह	वह
12	के	है	में	की	से	और	का	को	हैं	पर
13	एक	आप	और	यह	कर	हम	वह	पर	इस	अब
14	अत	अपना	अपनी	अपने	अभी	अंदर	आदि	आप	इत्यादि	इन
15	अंदर	अत	अदि	अप	अपना	अपनि	अपनी	अपने	अभि	अभी
16	पर	इन	वह	यिह	वुह	जिन्हें	जिन्हों	तिन्हें	तिन्हों	किन्हों
17	मैं	मुझको	मेरा	अपने आप को	हमने	हमारा	अपना	हम	आप	आपका
18	के	का	एक	में	की	है	यह	और	से	हैं
19	और	पर	एक	रत	कर	इस	यह	अन	वर	सम

To generate the list of the top ten stop words, the researcher used content from the Ministry of Electronics and Information Technology[1], the IIT Bombay's Centre for Indian Language Technology[2], Wikipedia dump[3], the open parallel corpus[4], a collection of subtitles, and the aesthetics text corpus that the researcher created using publicly available content. Owing to the fact that neither the publicly available stop word lists themselves, nor the methodology used to create them, were made available as official resources, the researcher was unable to access the corpora that were utilised to produce the publicly accessible lists.

RQ2: How can the available resources be used to compile a comprehensive list of stop lemma?

The observation made while trying to find the solution to RQ1 led the researcher to the use of the lemmas. In an effort to answer RQ2, the researcher assembled the top 100 stop words from 11 publicly accessible lists, combined all of the lists, and then substituted the words with the corresponding lemmas. This collection was termed as Set A. The researcher then created Set B that comprised a combination of the stop lemmas from 8 corpora. Set C was a comprehensive stop word list, that was generated by taking the intersection of Sets A and B.

Set A = $\cup_{i=1}^{11}$ generate_lemma(Set_i)

Here, Set_i refers to the set of stop words in source i, where 1<=i<=11, and generate_lemma is a function that substitutes all the words in the set with their corresponding lemmas.

Set B = $\cup_{i=12}^{19}$ Set_j

Here, Set_j refers to the set of the top 100 lemmas that have the highest frequencies in source j, where 12<=j<=19.

The values of i and j refer to the position of a language resource listed in Appendix H.

Set C = Set A ∩ Set B

The researcher reduced 1,096 words from publicly accessible lists to 1,071 words after the redundant phrases and words were eliminated. The distinctive lemmas of the words were compiled into a list. This list was 370 items long.

[1] https://www.cfilt.iitb.ac.in/

[2] https://tdil-dc.in/index.php/

[3] https://dumps.wikimedia.org/enwiki/

[4] https://opus.nlpl.eu/

There were 2,13,554 lemmas in Set B, which the researcher gathered by retrieving the lemmas for each word in the corpora. These lemmas were produced from 1,38,11,781 words, which were reduced to 4,05,111 words after duplicates were eliminated. Merging both lists and removing the common lemmas produced the final list of stop lemmas.

Table 3.3 shows that there are 311 lemmas in the list. From top to bottom, the lemmas are organised in decreasing order of frequency.

Table 3.3: List of stop lemmas

Stop Lemmas							
का	लोग	सारा	भाषा	अंत	छत	एलन	वग़ैरह
है	मिल	प्रकार	बदल	उधर	थक	किर	रक
वह	या	बडा	नीचे	कब	दोपहर	जर	वुह
में	फिर	नया	बंद	कुल	तहत	डल	
कर	वाला	निकल	मर	जबकि	पे	पडा	
हो	लिए	लिख	काफी	संख्या	ओह	रत	
यह	लेकिन	पानी	खुद	एस	तस्वीर	अल	
जा	तरह	इसलिए	बड़ा	ना	जादू	गर	
और	अब	एवं	पिता	हवा	बिंदु	चकमक	
से	बहुत	कई	शहर	परिवर्तन	सन	उम	
था	दिन	अभी	दुनिया	सर	शक	चन	
को	रख	अगर	विशेष	बहन	असल	दक	
मैं	जब	बाहर	जितना	सोना	धन्यवाद	नक	
नहीं	लगा	पूछ	उपयोग	वजह	एल	सकत	
पर	तथा	भारत	अंदर	मात्र	आह	बर	
रह	बाद	छोटा	खेल	मदद	लय	आत	
भी	चाह	सामने	लगभग	प्रकाश	सहमत	आद	
कि	यहाँ	बीच	स्वयं	खबर	पहल	ईस	
तो	दूसरा	तीन	अथवा	आग	नफरत	तया	
ले	घर	हर	मत	पद	मुझको	खक	
एक	समझ	जहाँ	पुरुष	अत	नरक	सबस	

Stop Lemmas							
दे	चाहिए	केवल	रास्ता	बह	दुबारा	रण	
ही	रूप	डाल	जैसे	आराम	गत	मक	
ने	जैसा	कितना	आवश्यकता	रस	सेट	करत	
अपना	पहले	बना	कल	दर	जनरल	यन	
जो	बार	वर्ष	भीतर	गलत	वर	उद	
आ	कभी	रात	कोशिश	खिलाफ	अप	साबुत	
कह	अच्छा	सबसे	प्रत्येक	जन	दोनो	अर	
कोई	बोल	कैसे	औरत	पालन	नह	यर	
हम	कारण	माँ	दस	आठ	आध	लन	
सक	ओर	बारे	खाना	जोड़	तर	षण	
आप	हाथ	आगे	दौरान	पल	कौनसा	एसे	
कुछ	कौन	भाग	वहीं	बिलकुल	कोन	क्यूंकि	
देख	आज	पी	सुबह	उच्च	लत	पत	
बात	पास	अलग	वर्ग	सदा	निहायत	वत	
साथ	पूरा	कहाँ	डर	अधिकांश	कवर	रद	
क्या	वहाँ	विकास	पढ	वन	बंदरगाह	टर	
दो	अधिक	प्राप्त	ए	निकट	दुसरा	यक	
तक	भर	कार्य	तुम्हारा	छह	तिन्ह	मह	
ऐसा	सुन	जगह	मा	आर	बाला	पनी	
लग	द्वारा	ऊपर	सच	ओ	तिस	उनकि	
चल	देश	शब्द	मतलब	पृथ्वी	उह	तथ	
सब	बता	बस	मानो	बज	तिसे	उनक	
बन	क्यों	ज्यादा	माध्यम	हल	यत	उत	

The complete list of stop lemmas resides at http://github.com/gayatrivenugopal/hindi-corpus-stoplemmas. It has been released under the GNU GPL v3 open source license. The researcher decided to make all the data and resources that were created as part of the research open source, so that other researchers can leverage the resources and extend this research, or use the data and resources in their own resource, without re-inventing the wheel. The researcher also aimed to make contributions in the area of Hindi language resources.

The process of creating the list was complete. However, it was essential to evaluate it by comparing it with an existing established resource. The researcher compared the list to the English stop word list offered by the Natural Language ToolKit (NLTK) package in Python, in order to evaluate this list. The Hindi words for the English stop words were translated using Google Translate[1] with manual interaction. Since the researcher could not verify whether the Hindi translations of other languages were accurate, the researcher chose English. The lemmas of these Hindi translations were created, and they were compared to the exhaustive stop lemma list. The results can be seen in the subsequent chapter.

3.3 Dataset

The dataset is a crucial resource that was created and used in this study. Since the researcher attempted to solve the problem of complex word identification as a supervised machine learning challenge, the study required a labelled dataset. The researcher conducted an annotation task in order to create labelled data for the research. An annotation task in natural language processing is the manual or semi-automated process of labelling or tagging a text or a given dataset. Annotations are essential for training and assessing machine learning models, because they generally offer further details about the features of the given text. Several natural language applications involve annotation studies that are essential for building and improving models for tasks such as named entity recognition, sentiment analysis, part-of-speech tagging, and more. To study the annotations and annotator preferences, the researcher aimed to determine the answers to the following questions:

RQ3: Is there a difference between annotators' native language and the language they felt most at ease reading?
RQ4: Is there a similarity between an annotator's preferred language and the official language of the area where they spent the most time?
RQ5: Is a word's lemma regarded as being simpler than a morphological variation of the word?
RQ6: Are the values of the features of words classified as complex and those classified as simple significantly different from one another?

[1]https//translate.google.co.in/

RQ7: Is there a difference in how well the model performs with respect to test data generated using annotations obtained from user categories formed using four criteria:

- Native language
- Hindi being the language that they were most comfortable with
- Years of formal training in Hindi
- Gender as specified by the annotator

Note that RQ7 tests the dataset against multiple biases.

The following sub-sections describe methods followed to reach the answers to the research questions RQ3 – RQ7.

3.3.1 Annotators

An important first goal that marked the beginning of the research was to conduct an annotation study with the overall goal of creating a carefully labelled dataset. This annotated dataset would act as the cornerstone for further analysis and offer insightful information about the understanding of words by annotators. Initially, the researcher attempted to obtain annotations by asking volunteers to participate in an annotation task, where given a group of sentences, they were asked to underline a word they could not understand. However, the results of the task indicated that the annotators did not read the sentences carefully, as many of the words that they selected as complex were in fact stop words. The researcher then set out to obtain funds for the study so that they can recruit annotators, and also create an online system for the study.

The annotation study that was designed after the pilot study failure, was sponsored by Symbiosis International (Deemed University). The study received a minor research project grant of 1,50,000 INR. The study was also granted approval by the Independent Ethics Committee of Symbiosis International (Deemed University) after two rounds of evaluation. The researcher recruited a group of 100 annotators by choosing residents of India who had completed formal education in Hindi in school. The researcher attempted to make the demographic distribution of the annotator cohort diverse. The study was enhanced by the inclusion of a diverse range of participants, including both working professionals and students, who were between the ages of 18 and 30. The median age of the annotators, was found to be 19.66. The standard deviation, which represents the dispersion within the age distribution, was

found to be 2.775. With 57 annotators self-identifying as male and 43 as female, gender representation was also taken into consideration, providing a fair perspective for each gender within the cohort.

The researcher chose not to divide the annotators according to language proficiency, as the requirement did not include being an expert in the language. This choice was based on the main objective of the research, which was to understand the complex dynamics of word comprehension in relation to different degrees of linguistic exposure. Understanding how people struggled with word comprehension, regardless of their level of language proficiency, was emphasised. The researcher used a methodical grouping technique in order to delve further into the complex effects of language exposure. The annotators were organised into 20 groups, each with 5 members, categorised according to their native language. Groups 1 through 10 were composed of non-native annotators whereas Groups 11 through 20 included annotators who were native speakers of the language. This composition was designed in order to conduct a comparison of native and non-native groups. To provide more depth and granularity to the subsequent analysis and findings of the research study, this intentional categorization sought to highlight potential differences in word comprehension based on differing levels of exposure to, and familiarity with the Hindi language. This categorisation would also be used for testing the dataset against biases.

3.3.2 Data for Annotation

The researcher used the aesthetics corpus that was created for the study. The researcher chose random sampling to choose sentences from the corpus. However, sentences with less than 5 words were removed. The researcher created 20 sets wherein each set contained 100 sentences. Each group of annotators received a different set of sentences. All the sets had 10 common sentences that were extracted from Twitter. The common sentences would be used to calculate inter-annotator agreement of the annotators when testing the dataset against biases. The researcher excluded writings from other fields such as history, law, technology, etc. as the goal was to investigate the criteria for word simplification in literature such as novels, short stories, and biographies.

3.3.3 Annotation Tasks

The researcher, with the help of a Computer Applications under-graduate student, designed and developed an online platform for the tasks including annotation. The platform was developed using JavaScript, and Flask, a Python-based framework. There were two tasks in the annotation process. The researcher designed the first task, Task 1, to identify those words for which the annotator was unable to comprehend the meanings of. The researcher designed the second task, Task 2, to assess the difficulty of the word, as well as its synonyms. The researcher created Task 2 with the goal of determining the difficulty of the target word, so that it could be compared to that of any alternative replacements, that is, the synonyms of the word. Each participant in Task 1 could annotate a maximum of 100 sentences. Each of the 5 annotators that belonged to a group received the same set of sentences, so that the researcher could gather words annotated by more than one annotator, in order to avoid bias. The annotators were asked to highlight any word in a sentence displayed on the screen, whose meaning they could not comprehend. The screens presented to the annotators can be seen in Figure 3.2 and Figure 3.3, respectively.

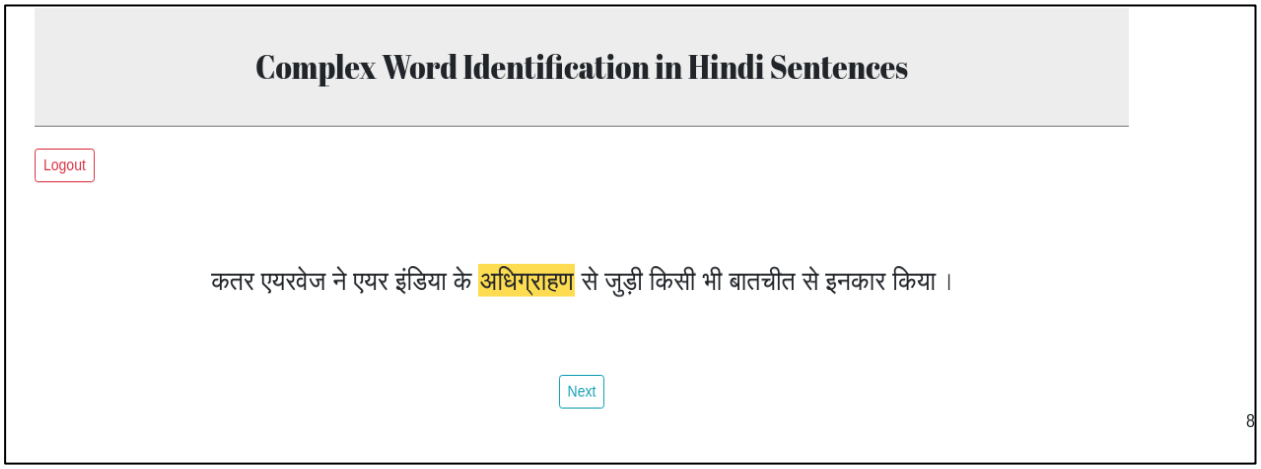

Figure 3.2: Example of a screen presented to the annotators in Task 1

The screen shows the result of highlighting a word that the annotator could not understand.

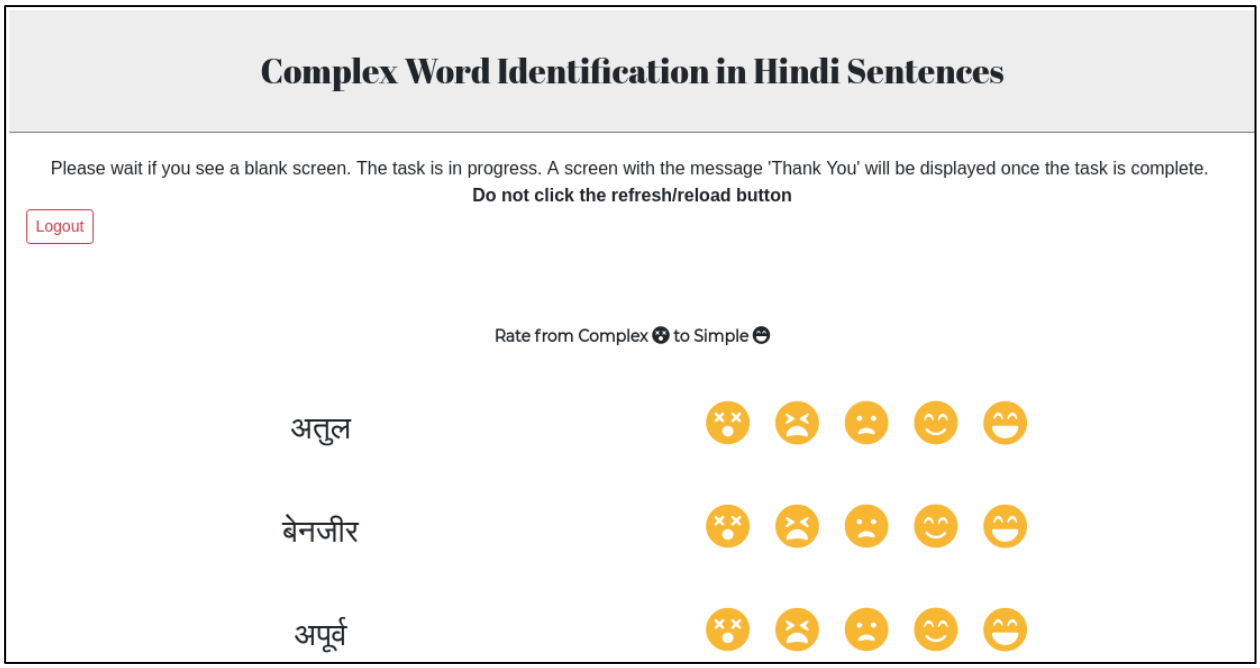

Figure 3.3: Example of a screen presented to the annotators in Task 2

In Task 1, the annotators were prohibited from choosing multi-word expressions with several words. The annotator would not indicate any word as complex if they were familiar with the meanings of the words. That is, it was not mandatory to annotate every sentence.

In Task 2, each annotator would see a set of words that included the lemma of the word that they had identified as complex in the first task, as well as its Hindi WordNet synonyms (Bhattacharyya, 2008). In Figure 3.3, the screen shows a list of synonyms of a word 'अतुल' that was highlighted as complex to understand, in Task 1. Each of these words was required to be rated by the annotators using an emoji. In order to prevent confusion and cognitive load in the case of several synonyms, the researcher did not employ relative rating, that is, ranking the difficulty level of a word and all its synonyms. The rating scale contained 5 emojis that ranged from depicting the angry emotion to the cheerful emotion, with angry denoting complex words and cheerful denoting simple words. The researcher gave clear instructions to the annotators before the annotation tasks, and was also available to answer queries during the entire duration of the tasks. A word was deemed complex if it received a rating of 3 or less. The researcher decided upon the number 3 as it was the median in the rating scale. However, since the number 3 was not associated with a happy emoji, a word that received a rating of 3 was also considered to be complex. The researcher included this task since it made more sense to compare a term with its synonyms than with unrelated words.

Figure 3.4 depicts the distribution of the annotations that Task 1 generated. The number of words that received agreement from all 5 annotators in a group was 109, that is, less than 5% of the 4,599 words that were annotated. In contrast, the count of words for which there was no agreement was 2,321, that is, almost 50%, thus demonstrating a low inter-annotator agreement. Figure 3.4 shows the distribution of 4,599 words in Task 1 and the count of annotators who were in agreement that a word was complex.

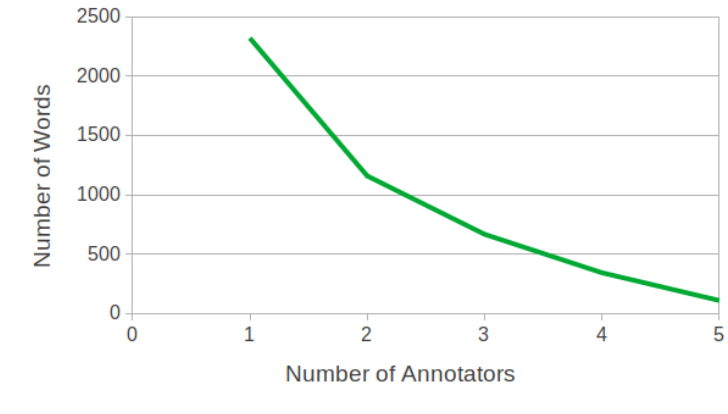

Figure 3.4: Distribution of annotations

Figure 3.5 shows the average number of annotations, i.e., the average of the number of words rated by annotators in each group.

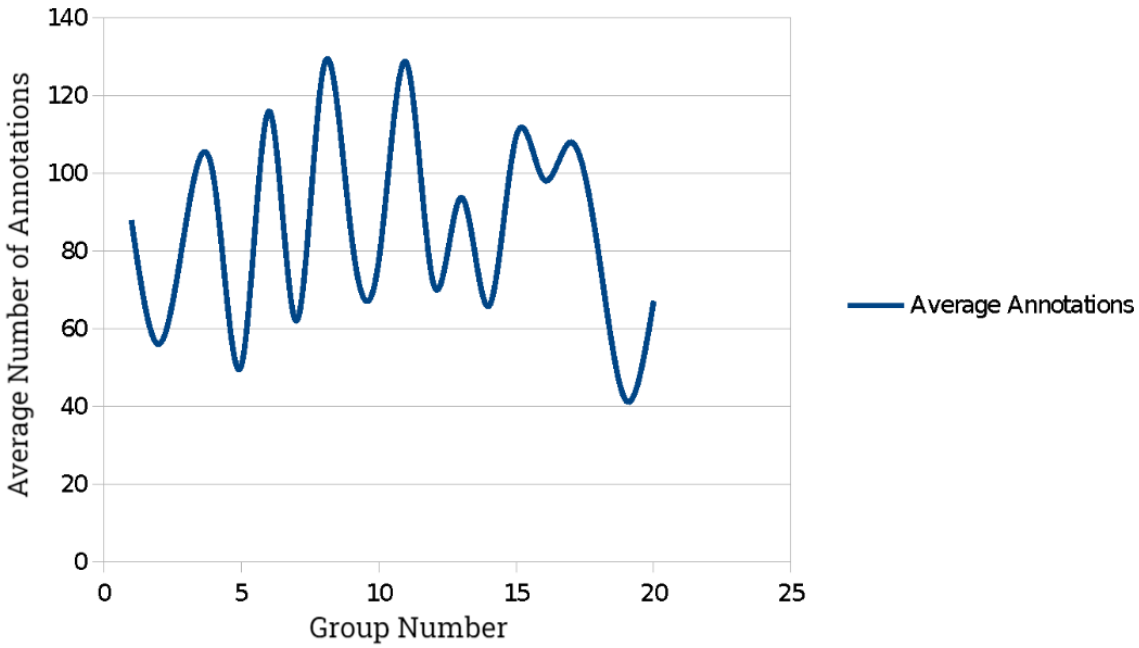

Figure 3.5: Average number of annotations in each group

The group of native speakers had the lowest number of annotations, but both types of groups had approximately the same number of annotations at their highest level. Consequently, the researcher was unable to limit the investigation to only non-native speakers. The data received from Task 2 is presented in Table 3.4. The annotation program displayed the word that the annotators selected as complex in Task 1, along with the synonyms of the word.

Table 3.4: Description of data obtained in Task 2

S.No.	Description	Count
1	Number of words ranked by participants	68,107
2	Total number of simple word annotations	35,471
3	Total number of complex word annotations	32,636
4	Total number of ranked words excluding digits and numbers	63,857
5	Number of unique words ranked by participants	18,186
6	Number of unique words annotated by at-least two participants	12,111
7	Number of unique words ranked by at-least two participants and that are present in the corpus	7,321

The annotators ranked each word in this list according to its level of complexity as perceived by them. The researcher cleaned the data by removing digits, as a few of the ranked words contained numbers. The researcher observed that in certain instances, the same word received several rankings from annotators. This may have occurred because of the varying contexts in which the word appeared. Since studying the complexity of a word based on its context was beyond the scope of this research, the researcher simply kept the first instance while removing the duplicate instances. The label to be assigned to a word was determined based on the average rank that it received. If only one annotator annotated a word, the researcher had to remove that word from consideration. Therefore, after the filtering process, the researcher chose words from the list of words that had been ranked by at least 2 annotators. The researcher only selected words that were included in the corpus because the dataset included corpus frequency as one of the features. If the word received an average rank of less than or equal to 3 in Task 2, the researcher assigned the label of 1 to the sample; otherwise, the researcher assigned it a label of 0.

The annotators ranked 68,107 words. 47.9% of words were assigned a rating greater than 3, which indicated that they were simple, while 52.1% of words were assigned a rating of 3 or less, which indicated that they were complex. The dataset contained 7,321 words after excluding numbers, redundant instances, and words that were not present in the corpus.

3.3.4 Classifier

Using a dataset of 5,111 records, the researcher carefully developed an ensemble classifier with the goal of improving the comprehension of lexical complexity. This dataset, which had 40% complex words and 60% simple words, aimed to capture the subtle differences in linguistic complexity. Owing to the training data, the classifier would be able to leverage the strength of a group of classifiers. The classifier used soft voting and a variety of models, including gradient boosting, extra trees, XGBoost, and random forest, to obtain its predictions. The soft voting method is an ensemble learning method that combines the predictions of different classifiers. For any given instance, each classifier generates its probability distribution over the classes. The probability distribution represents the probability of the instance belonging to each class. The predicted probabilities are averaged for each class. The class with the maximum averaged probability is selected as the final prediction. The equation associated with the soft voting method is as follows:

$\hat{y} = \text{argmax}_i \sum_{j=1}^{M} w_i p_{ij}$, where $\hat{y}$ is the predicted class, M is the count of models in the ensemble, w_j is the weight assigned to the prediction of the j^{th} model, and p_{ij} is the probability that is predicted by the j^{th} model for class i (Mohammed & Kora, 2022). The researcher used scikit-learn's soft voting classifier wherein the weights for each classifier are proportional to the performance of the classifier during the training stage.

A key component of the research study was the generation of several test sets, each purposefully constructed to examine the classifier's performance in various linguistic contexts. Task 2, which consisted of words ranked by annotator groups based on gender, linguistic skill, comfort level with Hindi, and academic training, formed the basis for these test sets. The evaluation process was enhanced by the intentional inclusion of male and female annotators, native and non-native speakers, and people with different degrees of comfort speaking Hindi or other languages.

Examining the complex relationship between lexical complexity and annotator traits, the researcher carried out a detailed analysis taking into account variables including academic training levels, linguistic comfort, and language exposure. To provide a more detailed analysis, the formal schooling years, which spanned from 1 to 16, were converted into a binary variable. The researcher separated those with substantial academic training (years higher than 8), and those with more modest educational backgrounds (years less than 9), by classifying the schooling years into high and low, with the threshold set at the median, that is, 8. This tactical choice not only allowed for a more targeted study, but it also supported the larger goal of identifying the complex connections between lexical complexity and annotator traits in the context of formal education.

3.4 Labelling Strategies

To classify a word as being 1 (complex), the researcher employed three methods. Following is a description of the methods used to label the dataset, and a list of the datasets created using those approaches.

3.4.1 Approach 1

A word was considered complex in the first approach if at least 2 annotators identified it as such in Task 2. There were 11,565 records in the dataset, with 9,382 simple labels and 2,183 complex labels. The training set and the test set were created using the k-fold cross-validation method with 5 splits, dividing the data into a 70:30 ratio for training and testing. 8,095 records made up the training set distribution, of which 18.88% were classified as complex (label 1) and 81.12% as simple (label 0). Due to the imbalance, a resampled dataset was used for training. For feature extraction, both undersampled and oversampled data were employed. Synthetic Minority Over-sampling Technique (SMOTE) was used for oversampling, whereas NearMiss-3 was employed for undersampling. By oversampling the minority class with artificial instances, SMOTE attempts to balance out datasets that are unbalanced. To even out the distribution of classes, NearMiss, on the other hand, focuses on removing specific examples from the majority class.

3.4.2 Approach 2

In the second approach, if a word received an average rating of less than or equal to 3, the researcher classified it as complex. Only words that were rated by at-least 2 participants were taken into consideration in this approach. Thus, a dataset of 7,321 records was created, of which 2,956 records were classified as complex and 4,365 records as simple. In order to avoid bias, the researcher did not consider words that were ranked by a single annotator. As a result, the dataset was smaller than the one generated using Approach 1. The ratio of training to testing was 70:30. The training set had 5,111 records, of which 40% had the label 1 and 60% records had the label 0. Resampling methods were not applied because this proportion was not undesirable.

3.4.3 Approach 3

To determine the agreement for each word, the researcher considered utilising the observed percentage of agreement. Since the researcher did not evaluate the agreement for a single item, the common inter-annotator agreement methods were inapplicable. Only words with observed agreement levels of more than 75% were included in the selection. As a result of the diverse backgrounds of the participants, there was a significant degree of loss of data because the agreement was not high. Therefore, the researcher labelled the word based on the majority vote. The dataset produced by Approach 3 contained 10,499 records. The researcher was left with 8,576 records after removing the instances where there was a tie. The researcher eliminated the words from the Hindi WordNet (Bhattacharyya, 2008) whose information was missing. Samples wherein the word was annotated by a single annotator were not taken into consideration. The resultant dataset contained 6,154 records, including 3,260 records with complex words and 2,894 records with simple words. The ratio of training data to testing data was 70:30, similar to the previous approach. The training set included 4,308 records, of which 47% were classified as simple and 53% as complex. Since the dataset was balanced, it did not require resampling.

3.5 Model Creation and Selection

The researcher treated the problem as a task requiring binary classification. The researcher extracted the lexical features and word embeddings of the words annotated by the participants to study the relationship between the features and the labels. Then, in order to select the best strategy, the researcher compared many classifiers and datasets labelled using various approaches.

3.5.1 Synset-Based Feature Normalization

Each group of words that the researcher constructed, represented a synset. A synset, which stands for synonym set, refers to a group of words that have similar meanings or are semantically connected. A synset is a collection of terms that are synonymous, or have comparable meanings. For example, consider the synset for the words - happy, joyful, and content. In this synset, these words are grouped together because they share a similar meaning related to positive or happy emotions. MinMax normalisation, which converts feature values into a consistent range between 0 and 1, was used by the researcher to guarantee a standardised comparison. Rather than applying the normalisation to the entire dataset, it was done separately on each synset of the word in the dataset. In order to generate normalised values, only related words, i.e., words that are a part of the same synset, were compared. It would not be fair to train a model by normalising based on feature values of unrelated words, such as "table" and "apple," which are words that belong to separate synsets. Thus, the researcher enabled a fair comparison of the feature values of words in various synsets by using synset-based normalisation. One of the key contributions of this study is Hindi complex word identification dataset created using synset-based normalized values.

3.5.2 Training and Test Sets

The counts of records with label numbers 0 (simple) and 1 (complex) were 4,365 and 2,956, respectively. Using 5,111 records as a training set, the researcher assigned labels of 1 and 0 to 2,044 records 3,067, respectively. The total number of records in the test set was 2,210, with 1,290 records labelled 0, and 920 records labelled 1.

3.5.3 Lexical Parameters of Classical Readability Formulae

A readability formula is a quantitative formula used to evaluate how easy or difficult it is to read a given text. To determine the degree of difficulty or grade level at which the text is understandable, these formulae usually take into account a variety of linguistic and structural elements of the text. Providing information about the accessibility of a piece of content for its target readership is the main objective of readability formulae. The researcher set out to explore the parameters of established mathematical readability formulae in various languages, and the features of words present in simple word lists, to determine the most suitable lexical features for the classifier.

Hindi WordNet (Bhattacharyya, 2008) was used to retrieve the values of the word's features. The researcher computed frequency from the corpora the researcher collated as well as created. The researcher started by establishing the relationship between the complexity and the various word properties. However, correlation can be deceptive because a given feature may not be significantly correlated with the target, but a group of features may be. To determine the important features, the researcher employed exhaustive feature selection and permutation feature importance.

Initially, the classifiers were trained using the default parameters. Permutation feature importance and exhaustive feature selection were used to calculate the feature importance values. The permutation feature importance, rather than the feature importance, for each model was determined. It describes the variation in the score of the model as a result of individually randomising each feature. The appropriate feature set for each model was determined using exhaustive feature selection. Exhaustive feature selection is a brute-force method that guarantees a full investigation of the whole feature space. It can, however, be computationally expensive, particularly for datasets containing a lot of features, because the number of combinations increases exponentially with the count of features. However, since the features were not very large, the researcher decided to proceed with this method. Accuracy and macro-F1 were employed as the evaluation criteria. The researcher calculated the value for the permutation feature importance for a certain feature by calculating the average of the permutation importance values over all folds in the 5-fold cross-validated model. The feature importance values were created by taking an average of the permutation importance values for features from undersampled and oversampled data in case of an imbalanced dataset. The

researcher created an intersection of the values by using the features produced by exhaustive feature selection from the oversampled data and the undersampled data, since exhaustive feature selection does not give as output, a set of continuous values.

As the next step, to find the most important features, the researcher employed the second strategy, which involved soft voting classification and random search model hyperparameter tuning. Soft voting pools the predictions from all the classifiers – support vector classifier, random forest classifier, nearest centroid classifier, extra trees classifier, random forest classifier, AdaBoost classifier, XGBoost classifier, gradient boosting classifier, and decision tree classifier. Each classifier develops its probability distribution over the classes for any given occurrence. The likelihood that an instance belongs to a particular class is represented by the probability distribution. For every class, the expected probabilities are averaged. As the final prediction, the class that obtained the highest averaged probability is chosen:
$\hat{y} = \text{argmax}_i \sum_{j=1}^{M} w_i p_{ij}$, where $\hat{y}$ is the predicted class, M is the count of models in the ensemble, that is, 9, w_j is the weight assigned to the prediction of the j^{th} model, and p_{ij} is the probability that is predicted by the j^{th} model for class i, where i is 0 or 1 (Mohammed & Kora, 2022).

Models were tuned using Receiver Operating Characteristic - Area Under the Curve (ROC AUC) scores. The balance between true positive rate, and the false positive rate for different thresholds in a binary classification model is represented graphically by the ROC curve. It facilitates the visualisation of the output of the model across different classification thresholds. The Area under the ROC curve is measured by AUC. It summarises the capability of the model to differentiate between the positive classes and the negative classes across a range of threshold values, which indicates the overall performance of the model. Better model performance is indicated by a higher AUC. Due to imbalanced datasets, Approach 1 employed precision-recall curves. The results were compared with ALL 1 and ALL 0 as the baselines. ALL 1 indicates that the prediction is always 1 for all the outputs, and ALL 0 indicates that the prediction is always 0 for all the outputs. Precision-recall curve is a machine learning tool for evaluating a classifier's performance, especially when there is an imbalance between the classes. The graph illustrates the precision vs recall trade-off at various probability thresholds. The curve indicates how the precision and recall values change as the threshold moves.

3.5.4 Word Embeddings as Features

Word representations in a continuous vector space, where semantically related words are mapped to adjacent points, are called pre-trained word embeddings. These embeddings are learned by training models on large text corpora and identifying the contextual relationships between words. This indicates that vector representations for words should be comparable if their meanings are similar. Predicting words in context is a common method in the training process, when the model understands the meaning of a target word by analysing the words that surround it. The researcher obtained the pre-trained word embeddings of the words from Kunchukuttan et al. (2020). These embeddings were trained on the AI4Bharat-IndicNLP dataset[1].

3.5.5 Classifiers and Evaluation Metrics

The researcher created four distinct datasets:

- a dataset containing lexical parameters, such as the length, number of synsets, synonyms, consonants, vowels, hypernyms, hyponyms, consonant conjuncts, number of syllables, and frequency of the words
- a dataset that exclusively contains pre-trained embeddings
- a dataset with pre-trained embeddings and word frequency
- a dataset with lexical characteristics, word frequency, and pre-trained embeddings

Based on the four datasets, the researcher was able to obtain four separate soft voting classifiers. An ablation test was also conducted to see if the classifier's performance would change significantly if any of the lexical features were eliminated. An ablation study is an experimental design that helps determine how different features or components affect the performance of a model. In machine learning, ablation refers to the removal of certain features. It entails methodically removing the features of a model in order to evaluate the effect on the overall performance of the model.

[1] https://github.com/AI4Bharat/indicnlp_corpus

Through the process of carefully deleting or modifying these features and monitoring the impact on the performance of the model, researchers can learn which features are essential to attaining favourable results. The ablation study evaluated each feature's importance and contribution to the prediction made by the model. The subsequent chapter contains a report on the findings from the ablation study.

Five models the researcher took into consideration for classification were ensemble models. In machine learning, an ensemble model is a method that combines the predictions of several individual models to obtain a prediction that is more reliable and accurate than any of the individual models by themselves. The theory behind ensemble learning is that individual model shortcomings can be compensated for, by integrating the strengths of other models, thus improving overall performance. Models including AdaBoost, decision tree, extra trees, gradient boosting, random forest, and XGBoost were used to classify the data. Here, decision tree is not an ensemble model. The classification method employed k-fold cross validation with 5 splits. Because of the observed collinearity between the features, the researcher decided against using logistic regression. It was also observed that ensemble models gave a better performance than a decision tree-based model. Therefore, the decision tree classifier was not considered in the subsequent experiments. The researcher first tuned the hyperparameters of the ensemble models randomly. The hyperparameters of the models were then tuned using grid search, and the researcher ultimately created a soft voting classifier as the final model. Hyperparameters refer to a model's configuration values in a machine learning problem. These must be set before training. Grid search is a method for hyperparameter tuning that includes a process of systematically going through a predetermined set of hyperparameter combinations for a particular model. Grid search analyses the performance of the model for every combination in a given grid to determine the ideal set of hyperparameters. Due to the inability to compute the metrics using AUC scores, the researcher could not tune the gradient boosting model and the extra trees model. Each of the three tuned and three non-tuned classifiers produced a probability distribution across the classes for every instance. Each class's probability of having an instance was represented by the probability distribution. Each class's projected odds were averaged. The final prediction was made for the class with the highest averaged likelihood:

$\hat{y} = \text{argmax}_i \sum_{j=1}^{M} w_i p_{ij}$, where $\hat{y}$ is the predicted class, M is the count of models, that is, 6, w_j is the weight assigned to the prediction of the j^{th} model, and p_{ij} is the probability that is predicted by the j^{th} model for class i, where i is 0 or 1 (Mohammed & Kora, 2022).

The classifiers and their corresponding hyperparameters can be seen in Table 3.5. The models were tuned using ROC scores as the metric.

Table 3.5: Hyperparameters used to train the classifiers

Classifier	**Hyperparameter**
Random Forest	Estimators, Maximum Depth, Minimum Samples Split, Minimum Samples Leaf, Maximum Features, Bootstrap
AdaBoost	Estimators, Learning Rate
XGBoost	Estimators, Learning Rate, Maximum Depth, Minimum Child Weight, Colsample_bytree

The description of each hyperparameter is as follows:

- Estimators – The count of trees or the number of boosting stages in an ensemble model
- Maximum Depth – The value of the maximum count of levels in a tree
- Minimum Samples Split – The value of the minimum count of samples needed to split an internal node
- Minimum Samples Leaf – The value of the minimum count of samples that a leaf node must have
- Maximum Features – The count of features to consider while searching for the best split
- Bootstrap – Determines whether the classifier used bootstrap samples while building trees
- Learning Rate – This value determines how fast the model must learn
- Minimum Child Weight – The value of the minimum sum of instance weight required in a child. This value is used to control overfitting
- Colsample_bytree – The value of the fraction of features to be randomly sampled for every tree

ALL 0, that is, a model that always predicted the outcome as 0, and ALL 1, that is, a model that always predicted the outcome as 1, served as baselines against which the data were compared. Additionally, the researcher carried out an ablation study, the findings of which are described in the subsequent chapter.

The researcher also attempted to use a neural network, although it was not planned to be used in the study. In order to create a neural network that can accurately predict complex words, the

researcher used word embeddings, frequency, and lexical data. ReLU activation functions were present at every layer of the neural network, which had three fully connected layers. The loss function that the researcher used was binary cross entropy.

To summarize, the following is a list of the methods and terms utilised in this study:

- Labelling Strategy
  - Approach 1: A word is classified as complex if at least two annotators gave it a score of three or lower.
  - Approach 2: If a word earned an average rating of 3 or less, it is considered complex.
  - Approach 3: If a word earned a majority rating of 3 or less, it is classified as complex.
- Models/Classifiers
  - Traditional Classifiers – Nearest Centroid, Decision Tree, and Support Vector Classifier
  - Ensemble Classifiers – Extra Trees, Random Forest, Gradient Boosting, AdaBoost, XGBoost, and Soft Voting Classifier
- Methods
  - Method 1: Values for feature importance were taken out of models that were trained using default hyperparameters
  - Method 2: A voting classifier was used to derive feature importance values from tuned ensemble models
- Features
  - Word embeddings and lexical features were used. The lexical features that were considered were count of synonyms, count of synsets, frequency, count of hyponyms, count of syllables, count of hypernyms, length, count of consonants, count of vowels, and count of consonant conjuncts.

3.6 Word Sense Disambiguation and Synonym Selection

The creation of a classifier for complex word identification formed the major part of the research. After creating a classifier, that is, after identifying the target word to be simplified, the next step is to identify the sense of the word, and select suitable simpler synonyms that could act as a replacement for the word in the sentence.

In their overview of SENSEVAL, an evaluation-based workshop that focuses on word sense disambiguation approaches, Edmonds & Cotton (2001) state that an all-words task must consist of at least 5000 words in running text, in which all the content words should be tagged. The researcher included 5,037 words of running text, with 2,067 unique words, for the study. The length of the sentences ranged from 6 to 46 tokens, with the average length being 18 tokens. The count of stop words was 2,980, out of which the count of unique stop lemmas was 442, whereas the count of unique stop words was 182. The count of content words was 2,057, out of which the count of unique content words was 1,625, whereas the count of unique content lemmas was 1,499. 596 records, that is, sentences were annotated by experts. Here, experts were teachers who taught Hindi or those who were actively involved in the editorial team of Hindi magazines. The observed inter-annotator agreement was 75%.

The researcher selected words that are present in BabelNet (Navigli & Ponzetto, 2020), and that have more than one synset, as disambiguation would not be required if there is only one synset for a word. BabelNet is a multilingual lexicalized semantic resource. It combines knowledge from various language resources such as WordNet, Wikipedia, and others, to create a unified representation of words and concepts across different languages. BabelNet provides a large-scale resource for multilingual natural language processing tasks, including word sense disambiguation, entity linking, and cross-lingual information retrieval. It offers rich information about word senses, translations, hypernyms, hyponyms, and other semantic relations, making it a valuable tool for researchers and developers working on multilingual applications.

The researcher used the BabelNet API[1] to retrieve synsets from the resource. The code was written using Google Colab[2], which is a hosted Jupyter Notebook service. The researcher pre-processed the text by removing punctuations including the symbol that refers to a period in Hindi ('|'), known as *poorna viram*. The researcher retrieved the context-sensitive embeddings of the target word, from indicBERT, mBERT, and XLNet language models, respectively.

IndicBERT is a language model specifically devised for natural language processing tasks in Indic languages (Joshi et al., 2020). It is based on the Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al, 2019), trained on huge quantities of text data from various Indic languages.

[1] https://babelnet.org/guide

[2] https://colab.research.google.com

IndicBERT captures the contextual information and semantic representations of words and sentences, enabling it to perform various downstream natural language processing tasks in Indic languages. mBERT, short for multilingual BERT, is a pretrained language model that is designed to handle multiple languages and perform various natural language processing tasks. The mBERT model extends the BERT architecture to handle multilingual tasks by jointly training on a diverse range of languages. It leverages the pretraining process to learn universal language representations that can capture contextual information and semantic relationships across different languages.

The researcher calculated the similarity of these embeddings with ARES sense embeddings (Scarlini et al., 2020b) and selected the sense with the maximum similarity. Scarlini et al. (2020b) created ARES sense embeddings that combined different contextual embeddings, such as ELMo (Peters et al., 2018) and BERT, to capture rich contextual information for disambiguating word senses. The framework leveraged large-scale sense-annotated corpora and pre-trained language models. The experimental results demonstrated that their approach outperformed existing state-of-the-art methods on multiple word sense disambiguation benchmarks.

The researcher created a script to generate synsets using the three language models for all the tokens in each sentence, in descending order of relevance, and stored them in three different files. Each file was associated with one language model. In order to select a model, the researcher compared the senses generated from each model, with the senses annotated by experts in Hindi. The researcher found that the senses generated using mBERT embeddings were closest to that annotated by experts, as compared to the senses generated using the other models. Although the accuracy of 48% was not high, when the researcher analysed the differences, the researcher noticed that there were senses in which the target word was repeated twice or thrice, whereas one sense consisted of just the target word. In this case, if the annotator selected the latter option and the model generated the former sense as the most suitable sense, the script would not consider them to be the same. For example, two of the synsets that was retrieved was the following:

Synset 1: ‘उद्योग', 'उद्योग', 'उद्योग', 'उद्योग', 'औद्योगिक', 'औद्योगिकीकरण', 'उद्योग', 'उद्योग', 'उद्योगों', 'औद्योगिक_विकास'

Synset 2: ‘उद्योग'

As can be seen, both synsets contain the same words. Here, if the model generated Synset 1 as its prediction, and the expert selected Synset 2, it would be considered as a mismatch of predictions, although, in reality it is not, in the given context. Therefore, the researcher did not work on improving the accuracy manually, as the objective of the researcher was to determine the model with the highest accuracy. The researcher observed a significant difference between the accuracy of mBERT, when compared with the accuracies of the other models. The researcher selected synonyms of the target word from BabelNet using the senses generated using mBERT embeddings, and leveraged the complex word identification model to determine the complexity of each synonym. The researcher used the model's probability of prediction as the factor to determine the best suitable replacement for the target word. In machine learning, a model's probability prediction is the likelihood or confidence that the model assigns to a specific outcome. Many machine learning models, notably those for classification tasks, produce probabilities rather than merely a binary prediction (class 0 or class 1). The likelihood that a certain input belongs to a particular class is estimated by the model and is represented by these probabilities. The synonym was presented to the user of the system only if the prediction's probability of the word being complex was less than that of the target word. If the probability was lower than that of the target word, no suitable alternative was proposed to the user.

Chapter 4

Results and Discussion

The sections in this chapter correspond to the sections of the Methodology chapter. Each section reports the results of the method adopted in the section.

4.1 Stop Lemma List Creation

The researcher attempted to create a list of stop words that can be used by various natural language processing tasks, to identify words and lemmas that can be ignored during the task. To determine the difference in consistency among stop words and stop lemmas in stopword lists, the first research question was framed as follows:

RQ1: Is there consistency among the top ten stop words in all the lists of available stop words, regardless of the position of the words in the lists?

The null hypothesis was designed as follows:

H_0: There is consistency among the top ten stop words in all the stop word lists, regardless of the order in which the words are listed.

H_A: The top 10 stop words for every list of stop words are different, irrespective of the order in which the words appear in the lists.

The top 10 stop words from LR1 to LR5 are listed in Table 4.1, and the top 10 stop words from LR6 to LR19 are listed in Table 4.2.

Table 4.1: Top ten stop words from LR1 to LR5

Language Resource	**Stop Words**									
1	है	के	में	की	और	से	का	को	नहीं	तो
2	है	के	मैं	नहीं	में	हैं	एक	आप	और	लिए
3	के	है	में	की	से	और	का	को	हैं	पर
4	के	में	की	और	लिए	हैं	है	से	का	को
5	के	है	में	और	की	से	एक	का	हैं	को

Table 4.2: Top ten stop words from LR6 to LR19

Language Resource	Stop Words									
6	के	है	में	की	का	से	और	को	हैं	भी
7	के	है	में	की	का	से	को	और	हैं	ने
8	है	के	में	से	की	को	का	हैं	और	हो
9	में	है	हैं	नहीं	लिए	गया	तथा	अपने	कुछ	साथ
10	जैसा	मैं	उसके	कि	वह	था	के लिए	पर	हैं	साथ
11	की	और	एक	तक	में	है	आप	कि	यह	वह
12	के	है	में	की	से	और	का	को	हैं	पर
13	एक	आप	और	यह	कर	हम	वह	पर	इस	अब
14	अत	अपना	अपनी	अपने	अभी	अंदर	आदि	आप	इत्यादि	इन
15	अंदर	अत	अलद	अप	अपना	अपनि	अपनी	अपने	अभि	अभी
16	पर	इन	वह	यही	वुह	जिन्हे	जिन्हो	तिन्हे	तिन्हो	किन्हो
17	मैं	मुझको	मेरा	तुम्हारा	हमने	हमारा	अपना	हम	आप	आपका
18	के	का	एक	में	की	है	यह	और	से	हैं
19	और	पर	एक	रत	कर	इस	यह	अन	वर	सम

Out of the 82 different stop words from the 19 available sources, the most sources where a word appeared, was 14 sources. Figure 4.1 displays a word cloud that was created using the list of these stop words and their frequency in the lists from all the sources.

Figure 4.1: Word cloud created using the top 10 stop words from 19 sources and based on their frequency

In the figure, the word count is proportional to the size of the word. The word lists throughout the sources were found to be inconsistent. Out of the 82 different stop words from the 19 sources, the most sources where a word was present, was 14 sources.

The researcher also discovered inconsistencies in the list of lemmas. Table 4.3 displays the top 10 stop lemmas from each source.

Table 4.3: Top ten stop lemmas from each source

Language Resource	**Stop Lemmas**									
1	का	है	वह	हो	में	कर	था	जा	यह	और
2	है	का	मैं	कर	वह	यह	नहीं	हो	में	आप
3	का	है	मैं	वह	हो	कर	यह	जा	से	और
4	का	है	मैं	और	कर	जा	से	को	दे	यह
5	का	है	कर	में	और	से	एक	जा	यह	ले
6	का	है	कर	में	यह	हो	जा	से	और	वह
7	का	है	कर	में	यह	हो	वह	से	को	जा
8	है	का	कर	में	हो	से	यह	जा	को	और

Table 4.2 has fewer language resources than Table 4.1 and Table 4.2, as there were only two stop lemma lists that were publicly available. As a result, the researcher created the lemmas of the terms using the corpora the researcher already had, and arranged them according to their raw frequencies. In Figure 2, a word cloud that was created using the top 10 lemmas from 8 different sources (LR- 1 to LR-8) is displayed.

Figure 4.2: Word cloud created using the top 10 stop lemmas from 8 sources

There were 22 distinct stop lemmas among the 8 sources utilised in the study, and the most sources where a lemma was present, was 8. This is more than the count of words in the stop word list. In light of this, the researcher believes that when constructing a thorough list of stop words, the researcher needs to take into account the lemmas of the words rather than only their numerous morphological forms.The researcher rejected the null hypothesis and accepted the alternative hypothesis.

Significant implications for text analysis and natural language processing (NLP) result from the examination of the consistency between stop words and stop lemmas in this study. The inconsistency of stop words raises questions about the existence of a common set of stop words. This result suggests that lemmatization does not completely reduce variability, even though it does lead to a more consistent representation when compared to individual morphological forms. Nonetheless, the comparatively better consistency seen in lemmas as opposed to words implies that taking stop lemmas into account can be a useful tactic for creating stop word lists that are more robust. The findings and subsequent inherent suggestion that when creating stop word lists, lemmas should be given precedence over morphological forms of words has wider implications for natural language processing applications. This emphasises the importance of considering the underlying semantics and linguistic structure in addition to surface-level word forms. Researchers and practitioners could use this information to help them refine their stop word selection strategies, which will result in more efficient and domain-specific natural language processing models. As a result, the work advances text processing methods for natural language understanding by drawing attention to the necessity of continued stop word list customisation and refining, depending on the linguistic nuances of specific datasets or applications.

The second research question dealt with the process of creating and evaluating a comprehensive stop lemma list. In the methodology chapter, the researcher explained the process of the creation of the stop lemma list. In this chapter, the researcher describes the evaluation of the generated stop lemma list.

RQ2: How can the available resources be used to compile a comprehensive list of stop lemma?

It was crucial to assess the generated stop lemma list against a known and current reference. The researcher evaluated this list by comparing it to the list of English stop words provided by the Natural Language ToolKit (NLTK) package in Python. The researcher studied the Hindi translations of the English stop words. These translations into Hindi were made, and their lemmas were checked against the comprehensive stop lemma list. The 179 words from the English stop word list were converted into 74 distinct equivalent Hindi stop lemmas. The terms "being," "will," and "shall" were not translatable. The researcher expanded the word form "ll" to include "will" as well as "shall," even though they were not included as complete words in the English stop word list. By studying the terms in the English stop words list, an ambiguity of a specific kind was discovered. Along with words like "shan" and "aren," the word "win" was on the list. In this context, the researcher did not take the past tense of "win" into consideration when defining "won." It was a variant of wouldn't, that is, won't, in which the apostrophe and the letter "t" were dropped. For words like "shant," "aren," etc., the researcher used the same procedure. During the process of disambiguation, translation, and lemmatization, the researcher discovered that the stop lemma list contained 73 out of the 74 distinct lemmas. The lemma that was missing from the list was 'जरूर' which is translated to another stop word 'must' in English. Therefore, it is evident that the stop lemma list that the researcher created is similar in nature to the list of English words that are commonly used as stop words, and used widely by researchers using NLTK.

4.2 Dataset Creation

The observations made in the data collection study carried out in the context of the annotation study, as well as the effectiveness of a classifier on various dataset types, are included in this section. There are four subsections in this section:

- Analysis of annotation jobs
- Dataset that includes a study of the values of the dataset features
- Classifier Evaluation, which is composed of the outcomes of analyses carried out on specialised subsets of the dataset.
- Further Observations

In each sub-section, the researcher answers the research questions mentioned in the Methodology chapter.

4.2.1 Annotation Tasks

The study aimed to give a better understanding of the linguistic diversity, preferences, and perceived word complexity of a diverse set of readers. The following research questions and the answers obtained by the researcher, provide a better understanding of the complexity of this research.

RQ3: Is there a difference between annotators' native language and the language they felt most at ease reading?
72% of annotators felt more at ease reading literature written in a language other than their own. Because English is used frequently in schools, universities, and daily life. 93.05% of the annotators selected English as their preferred language for reading. Also, the researcher observed that annotators who were native Hindi speakers did not select Hindi as their preferred reading language.

RQ4: Is there a similarity between an annotator's preferred language and the official language of the area where they spent the most time?
66% of annotators selected English as the language they felt most at ease reading. However, 24% of annotators selected the official language of the residing region in which they spent the majority of their years, suggesting that a person's preference for a language may be influenced by the region in which they spend the majority of their time.

The researcher observed notable implications for language preferences and environmental influences by investigating possible differences between the native languages of annotators and the languages they felt most comfortable reading. The fact that 72% of annotators said they were more comfortable reading literature written in a language other than their native tongue, specifically, English, highlights how widely used English is in everyday life, academia, and education. The fact that a high percentage of annotators chose English as their favourite reading language is indicative of this prevalence. The finding that native Hindi speakers, did not select Hindi as their favourite reading language highlights the intricate dynamics of underlying language preferences, which may be impacted by variables other than linguistic nativity.

Further adding complexity to language choice is the investigation of the connection between an annotator's language of preference and the official language of the area in which they spent most of their time. The results suggest that linguistic choice and exposure to the environment may be related. This suggests that linguistic landscapes of the places people live can shape linguistic affinities of people, thus demonstrating the complex interaction between personal preference and external factors. These results have ramifications for a number of disciplines, including language policy, cultural studies, and education. By acknowledging the dominance of some languages and maybe promoting inclusivity for people from varied linguistic backgrounds, educational practices can be informed by an understanding of the variables that influence language preferences. The understanding that linguistic preferences can be influenced by environmental context further emphasises how crucial it is to take local linguistic dynamics into account when developing language policies and initiatives.

RQ5: Is a word's lemma regarded as being simpler than a morphological variation of the word? 8,744 words in Task 1 were classified as complex. 2,424 words were marked as complex in Task 1 where they were not in their respective lemmatized forms, but were marked as simple in Task 2, where they were in their respective lemmatized forms. With a low percentage of 27.72%, it is possible that the word's lemmatized form had little impact on how difficult readers thought it was.

In a context-free setting, the researcher discovered that a word's complexity does not change considerably when it is provided in its lemmatized form. The researcher cannot therefore presume that in such a situation, the morphological variant of a word should be regarded as distinct from the word itself. This challenges the notion that the morphological variant should be distinctly considered from the word itself. However, the researcher acknowledges the need for further investigation into the impact of morphological diversity in identifying complex words, especially in scenarios where context is crucial. This suggests avenues for future research to delve deeper into the interplay between morphological variations and contextual understanding in language processing tasks, providing insights that could enhance the effectiveness of natural language processing models.

Of the 50 native speakers, the researcher discovered that 70% specified a language other than Hindi as their preferred language. The preferred language of many annotators was English, which is known as the global lingua franca (House, 2014; Smokotin et al., 2014). Several

annotators were more at ease using English because Hindi or a non-Hindi native language is sometimes overshadowed by the usage of English, particularly in an academic atmosphere. Even though it was not their native language, about one-fourth of the annotators selected the official language of the residing region as the one they felt most at ease with. Consequently, despite the fact that the annotators were native Hindi speakers, the researcher could not presume that they were fluent in the language. A few of the annotators were also non-native Hindi speakers who had studied it for more than eight years.

More sentences were annotated by non-native speakers in comparison to the count of sentences that were annotated by native annotators. However, compared to non-native annotators, native annotators annotated on average more than twice as many sentences. The fact that the average number of annotations is the same for native and non-native speakers emphasises the subjective and challenging nature of the research problem. Figure 4.3 indicates a low level of inter-annotator agreement.

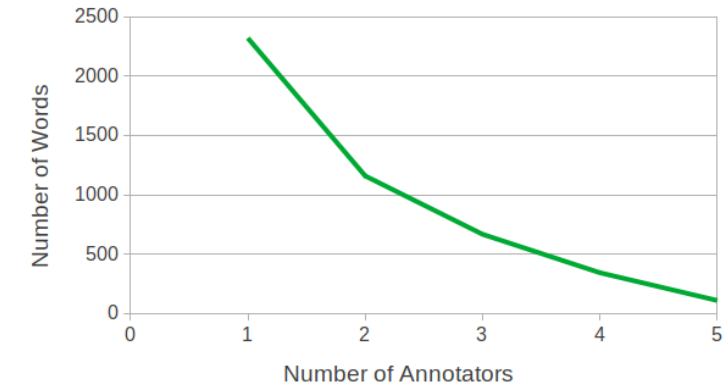

Figure 4.3: Distribution of 4,599 words and the annotators who agreed on a particular word being complex in Task 1

This was also observed in past studies where the non-native annotators came from a diversity of linguistic backgrounds leading to a low inter-annotator agreement value (Paetzold and Specia, 2016b). There was no connection between the total number of words that were annotated by an annotator and the number of years of academic training in Hindi they received. Yet, those who frequently read Hindi-language content annotated fewer words than those who did not. As a result, the researcher deduced that reading habits rather than academic instruction may be a key indicator of language familiarity.

The implications that arise from these findings are important for comprehending the behaviour of annotators in language studies. They provide insight into the complex aspects that affect annotation patterns. It is evident from the finding that native annotators annotated an average of more sentences than non-native speakers, who annotated fewer sentences overall. This underscores the subjectivity and difficulty of the annotation process. The fact that both native and non-native speakers annotated the same amount of content on average highlights the complexity of the subject, and raises the possibility that annotation difficulties may not be solely based on language proficiency. The researcher's conclusion that reading habits are a better measure of language familiarity than formal academic preparation is a useful finding for language evaluation and instruction.

4.2.2 Dataset

RQ6: Are the values of the features of words classified as complex and those classified as simple significantly different from one another?
The features of the words classified as complex and simple are shown in Table 4.4, along with their respective means and standard deviations (S.D.).

The researcher used minmax normalisation to create the dataset based on the feature values of the target word's senses. As a result, it is expected that the dataset would accurately reflect the situation in real life where a word is compared with related words rather than words that are unrelated. In order to determine whether there was a statistically significant difference between the values of each characteristic in the two different groups of simple words and complex words, the researcher chose to use the Mann-Whitney U test (Mann and Whitney, 1947). The decision was driven by the data's non-normal distribution, which demanded a reliable non-parametric test. After running the test and receiving an extremely low p-value of 0.00, the researcher was able to identify statistically significant differences in each of the variables that were being examined. The significance of these variations highlights the possibility of using the features in the dataset to successfully differentiate between simple and complex words. Significant ramifications arise from this finding, which implies that the features found and examined are representative of language complexity. Building on these discoveries, future studies could investigate the specific features that play a key role in this differentiation, providing a more

comprehensive knowledge of the linguistic elements that are essential in distinguishing between simple and complex words.

Table 4.4: Mean and standard deviation of the dataset's features for both complex and simple words

Features	Mean		Standard Deviation	
	Simple	Complex	Simple	Complex
Word Length	0.38	0.49	0.27	0.27
Count of Synsets	0.28	0.18	0.35	0.28
Count of Synonyms	0.33	0.26	0.32	0.29
Count of Consonants	0.38	0.49	0.3	0.29
Count of Vowels	0.43	0.52	0.29	0.29
Count of Hypernyms	0.26	0.52	0.33	0.29
Count of Hyponyms	0.23	0.15	0.34	0.29
Count of Consonant Conjuncts	0.23	0.31	0.35	0.37
Count of Syllables	0.40	0.50	0.31	0.30
Lemma Frequency	0.15	0.03	0.27	0.12

In order to compare the values of words classified as complex and simple, the researcher calculated the average values of the features. In line with findings from earlier studies, the researcher found that the values of complex word qualities, such as length, were marginally

higher than those of words classified as simple, and vice versa for frequency (Kauchak, 2016; Quijada and Medero, 2016).

4.2.3 Classifier Evaluation

RQ7: Is there a difference in how well the model performs with respect to test data generated using annotations obtained from user categories formed using the following criteria?

- Native language
- Hindi being the language that they were most comfortable with
- Years of academic training in Hindi
- Self-reported gender

Table 4.5 and Table 4.6 report the classifier's performance on a variety of specialised datasets.

Table 4.5: The classifier's performance on dataset types categorized by native language speakers and Hindi language preference, and the proportion of complex and simple terms in each dataset type

Type of Dataset	**AUC Score**	**F1 Score**	**% Complex Words**	**% Simple Words**
Non-Native Speakers	0.601	0.449	57.26	42.74
Native Speakers	0.668	0.528	55.88	44.12
Annotators most comfortable with Hindi	0.699	0.581	54.59	45.41
Annotators most comfortable with another language	0.650	0.548	54.27	45.73

Table 4.6: The classifier's performance on dataset types categorized by annotators with formal training in Hindi and self-reported gender, and the proportion of complex and simple terms in each dataset type

Type of Dataset	**AUC Score**	**F1 Score**	**% Complex Words**	**% Simple Words**
Annotators with low formal training	0.66	0.53	55.54	44.46
Annotators with high formal training	0.74	0.68	52.15	47.85
Annotators who identified as females	0.71	0.61	53.71	46.29
Annotators who identified as males	0.66	0.56	53.83	46.17

AUC and F1 scores were used to assess the model. The researcher observed that native speakers, annotators with a higher preference for Hindi as compared to other languages, annotators with strong academic qualifications, and annotators who identified as females, all received higher ratings, despite the model not being overly biased in favour of any particular annotator category. Despite equally splitting the annotators into native and non-native speakers, the researcher found that the native annotators' test set performed marginally better than the non-native annotators' test set.

For each of the common ten sentences that were assigned to all annotators, the agreement values were determined using Krippendorff's alpha (Krippendorff, 2011). Krippendorff's alpha is a metric for evaluating the consistency or reliability of coding or annotation carried out by several annotators. This metric, which was created by Klaus Krippendorff, is especially helpful when working with nominal, ordinal, or interval data. The range of the Krippendorff's alpha is 0 to 1, with 0 denoting no agreement that goes above and beyond what would be predicted by chance, and 1 indicating perfect agreement. When more than two annotators are involved and there may be missing data, researchers and analysts frequently use Krippendorff's alpha. Assessing inter-

rater reliability in a variety of research situations is made adaptable by its applicability to cases with varying levels of measurement.

The fact that the group of native speakers' annotators exhibited slightly higher inter-annotator agreement (0.193) than the group of non-native annotators (0.179) suggests a connection between these two findings. Based on this finding, the researcher assumed that the predictions made on the native annotators' test set are more accurate than those made on the non-native annotators' test set. Similar results were obtained for the other groups, which revealed that the annotators who selected Hindi as their favourite language had a marginally higher agreement coefficient (0.381), as compared to the annotators who did not (0.158). Accordingly, the agreement coefficient value of the highly trained Hindi annotators (0.207) was higher than that of the less highly trained Hindi annotators (0.145). However, despite the fact that the model performed better for annotations produced by annotators who self-identified as females (0.187 and 0.189, respectively), there was no statistically significant difference between the inter-annotator agreement scores of male and female annotators.

The implications drawn from this study have wide-reaching significance for both model developers and practitioners in natural language processing. The identification of potential biases is required and mitigation strategies in model training and evaluation must be planned. Future research endeavors should aim to delve deeper into these complexities, fostering a more comprehensive understanding of the different factors shaping the effectiveness and fairness of natural language processing models in diverse linguistic and demographic contexts.

4.2.4 Other Observations

Significant observations relating to annotations that were not covered in the preceding subsections are as follows:

- Using Krippendorff's alpha (Krippendorff, 2011), the inter-annotator agreement for the native and non-native groups was calculated to be 0.2421 and 0.1143, with an average of 0.1782. Given the diverse backgrounds and vocabularies of the annotators, it was anticipated that this value would be low.

- 5,213 sentences totaling 2,768 by non-native speakers and 2,445 by native readers were annotated. Native readers annotated, on average, 98.82 sentences, compared to 55.36 for non-native readers. This defies the notion that words would be easier to understand for native readers than for non-native readers.
- The total number of words that were annotated by native speakers was 3,911, compared to 4,645 by non-native speakers. Although the group with native annotators has the lowest number of annotations, both types of groups have an equivalent range of annotations. The range was 195 for the nonnative annotators, and 215 for native annotators. The average number of annotations generated by native and non-native annotators was 86.3 and 84.82 respectively, demonstrating that both groups had similar vocabulary restrictions. The total number of words annotated in task 1 did not correlate with the annotators' ages ($r = -0.121$).
- In comparison to male annotators, who made an average of 85.228 annotations, female annotators made an average of 86.233 annotations.
- Annotators who read in Hindi on a regular basis annotated 78.67 words on average, as opposed to 86.61 words on average for annotators who did not.
- The count of words in Task 1 that were annotated as complex did not significantly correlate with the number of years spent learning Hindi as a required subject in school ($r = 0.203$).

The calculated values of Krippendorff's alpha for the native and non-native annotator groups, as well as the predicted low value resulting from different backgrounds, highlight the intrinsic difficulty of reaching consensus in annotation jobs. This highlights the significance of recognising and resolving the possible subjectivity and diversity among annotators, requiring careful thought to be given to the interpretation of annotation findings and training of models for natural language processing tasks. The assumption that certain words could be simpler for native readers to understand, is called into question by the difference in the number of annotated sentences between native and non-native annotators. Even though the overall count of sentences differs between the two groups, the average number of annotations per annotator shows that the vocabulary limits are similar. This discovery highlights the complex nature of language comprehension by emphasising that linguistic difficulties and comprehension variations are not solely determined by the native or the non-native status of the annotator. The reading habits of annotators appear to influence their annotation preferences, as is evident from the difference in the average number of words annotated between those who read Hindi regularly and those who do not. This suggests that the capability of the annotators to understand and evaluate linguistic complexity may be influenced by their familiarity with the language in written form,

underscoring the significance of taking reading habits into account in annotation studies. This study creates opportunities for more research into the many factors that affect language assessments by indicating that factors other than formal schooling might influence annotator judgements.

The aim of this research was to devise a tool to recognize complex words in a given Hindi text. In order to create a dataset of words that had been classified as complex or simple based on lexical and semantic characteristics as well as the frequency of the word, the researcher initially set out to carry out an annotation task. Both native speakers and non-native annotators were recruited by the researcher. A native annotator is someone who learned the language of the country or area where they were born. The researcher found that the issue of complex word identification is difficult due to the varying amounts of exposure that people have to the language.

The researcher discovered that although regular reading can significantly increase vocabulary, years of academic language study may not be the only element in determining a person's familiarity with language words. The classifier's performance was tested utilising several specialised datasets, and the researcher also observed these variances in performance. The inclusion of the features in the dataset was justified by the researcher's observation that the values of the features varied significantly between simple and complicated terms.

4.3 Model Creation and Selection

This section describes the results of the methods undertaken during feature selection, and model creation and selection.

4.3.1 Lexical Parameters of Classical Readability Formulae

The objective of the study was to ascertain if it is possible to determine that a given word in a certain Hindi sentence is simple or complex, using the criteria of conventional readability formulae, which are used to grade the readability of both English and non-English language content. Using a variety of methods, the researcher attempted to identify the relevant features from eight models, five of which were ensemble models based on trees. A voting classifier was

also employed. Using the exhaustive feature selection method and the permutation feature importance approach, the researcher determined the feature importance using performance metrics such as accuracy and macro-F1 scores. Note that this sub-section focusses on determining the relevance of the lexical features used to create the classification model. The study intends to provide thorough insights into the linguistic properties that support complex word identification in Hindi sentences by exploring the nuances of feature selection and importance. In order to achieve the main objectives of the study, the investigation of several models and assessment metrics highlights the depth and comprehensiveness of the methodology used in exploring the intricacies of lexical feature importance.

The following is a list of the methods and terminology used in this experiment:

- Labeling Approach
  - Approach 1 – A word was deemed to be complex if a minimum of two annotators assigned a rating of 3 or less to the word
  - Approach 2 - A word was deemed to be complex if the average of the ratings assigned to it was 3 or less
  - Approach 3 – A word was deemed to be complex if the majority rating assigned to it was 3 or less
- Models
  - Traditional Classifiers – Support Vector Classifier, Nearest Centroid, Decision Tree
  - Ensemble Tree-Based Classifiers – Random Forest, Extra Trees, AdaBoost, XGBoost, Gradient Boosting, Soft Voting Classifier
- Methods
  - Method 1 – Models that were trained with default hyperparameters yielded feature importance values
  - Method 2 – Tuned ensemble models and a voting classifier yielded feature importance values.
- Features

The features that were considered were – length, count of synonyms, count of synsets, frequency, count of hyponyms, count of syllables, count of hypernyms, count of consonants, count of vowels, count of consonant conjuncts.

The researcher recorded the results of each of the eight models after the use of Method 1, with resampling for Approach 1, and without resampling for Approach 2 and Approach 3. Even

though the researcher initially only selected six features, the results for additional features that were used in the study based on the observations from the literature, are described. These features included the count of synonyms, synsets, hyponyms, and hypernyms. The highest rank of 1 was given to the feature with the highest average after the researcher calculated the feature importance values using accuracy and macro-F1 scores. With respect to exhaustive feature selection, a feature was assigned a value of 1 if it was included in a feature subset for a model, and a value of 0 if it was not included.

A feature subset was selected for each model created by using datasets labelled using each approach, with the help of an exhaustive feature subset selection process. The accuracy score and the macro-F1 score were used as the evaluation metrics. If a feature was a part of a feature subset for a model, it was assigned the value 1, otherwise it was assigned the value 0. The researcher separately computed the values obtained using the accuracy score and those calculated using the macro-F1 score, and also calculated the average of the values for each feature across all the models. Figure 4.4 displays the ROC curves of the models.

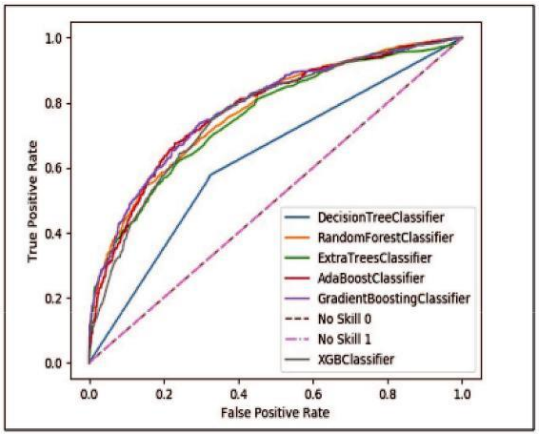

Figure 4.4: ROC curve for tree-based models

Table 4.7 displays the results for Approach 1.

Table 4.7: Feature importance values for each model created using Approach 1 that were calculated using accuracy and macro-F1 metrics

Feature	Permutation Feature Importance		Exhaustive Feature Selection	
	Accuracy	Macro-F1	Accuracy	Macro-F1
Count of synonyms	8.4	8.4	0.4	0.4
Count of synsets	7.5	7.6	0.7	0.5
Frequency	7.4	7.5	0.3	0.3
Count of hyponyms	7	7	0.4	0.4
Count of syllables	6.4	6.4	0.3	0.3
Count of hypernyms	6	5.8	0.4	0.4
Length	3.9	3.8	0.1	0.1
Count of consonants	3.4	3.3	0.4	0.4
Mean	5.5	5.5	0.3	0.3
Median	6.1	6.1	0.3	0.3
Standard Deviation	2.2	2.2	0.2	0.1

The results indicate that the accuracy and the macro-F1 scores are similar for the features. The researcher observed a similar pattern for the values calculated on the datasets created using Approach 2 and Approach 3. Accuracy is often used as a primary evaluation metric to gauge the overall performance of a machine learning model. In permutation feature importance, accuracy is calculated before and after permuting feature values to determine the impact of each feature on the model's accuracy. In both, permutation feature importance and exhaustive feature selection, the selected evaluation scores are essential for understanding the impact of individual

features on the model's overall performance and for guiding the feature selection process. Table 4.8 displays the results for Approach 2.

Table 4.8: Feature importance values for each model created using Approach 2 that were calculated using accuracy and macro-F1 metrics

Feature	Permutation Feature Importance		Exhaustive Feature	
	Accuracy	Macro-F1	Accuracy	Macro-F1
Count of synonyms	5.1	5.8	0	0
Count of synsets	4.8	4.8	0	0
Frequency	10	10	0.63	0.63
Count of hyponyms	8.3	8	0	0
Count of syllables	3	2.8	0	0
Count of hypernyms	4.8	5	0	0
Length	4.1	4.1	0	0
Count of consonants	6.5	6.5	0	0
Count of vowels	3.8	3.5	0	0
Count of consonant conjuncts	4.8	4.6	0	0
Mean	5.5	5.5	0.06	0.0
Median	4.8	4.9	0	0
Standard Deviation	2.2	2.2	0.2	0.2

The researcher observed that the features with the highest scores are different in Approach 1 and Approach 2. Table 4.9 displays the results for Approach 3. The values are similar to that of Approach 2.

Table 4.9: Feature importance values for each model created using Approach 3, calculated using accuracy score and macro-F1 metrics

Feature	Permutation Feature Importance		Exhaustive Feature Selection	
	Accuracy	Macro-F1	Accuracy	Macro-F1
Count of synonyms	5.1	5.8	0	0
Count of synsets	4.6	4.6	0	0

Feature	Permutation Feature Importance		Exhaustive Feature Selection	
	Accuracy	**Macro-F1**	**Accuracy**	**Macro-F1**
Frequency	10	10	0.6	0.6
Count of hyponyms	8.3	8	0	0
Count of syllables	3	2.8	0	0
Count of hypernyms	4.8	5	0	0
Length	4.1	4.1	0	0
Count of consonants	6.5	6.5	0	0
Count of vowels	3.8	3.5	0	0
Count of consonant conjuncts	4.9	4.8	0	0
Mean	5.5	5.5	0.1	0.1
Median	4.8	4.9	0	0
Standard Deviation	2.2	2.2	0.2	0.2

The researcher combined the results produced by the exhaustive feature subset and permutation feature importance methods based on accuracy and macro-F1 independently for each approach. The average of the accuracy values and macro-F1 values was calculated for each feature in each approach. The values for Approaches 1, 2, and 3 are presented in Table 4.10, Table 4.11, and Table 4.12, respectively.

Table 4.10: Feature importance values for every feature across all models using Approach 1

Feature	Importance Value based on		Aggregate
	Accuracy	Macro-F1	
Count of synonyms	8.8	8.8	8.8
Count of synsets	8.1	8.1	8.1
Frequency	7.6	7.8	7.7
Count of hyponyms	7.4	7.4	7.4
Count of syllables	6.6	6.6	6.6
Count of hypernyms	6.4	6.1	6.3
Length	4	3.9	4.0
Count of consonants	3.8	3.6	3.7
Count of vowels	3.1	3.3	3.2
Count of consonant conjuncts	2.4	2.5	2.4

Table 4.11: Aggregate of the feature importance values for every feature across all models for Approach 2

Feature	Importance Value based on		Aggregate
	Accuracy	Macro-F1	
Count of synonyms	5.1	5.8	5.4
Count of synsets	8.1	8.1	8.1
Frequency	10.6	10.6	10.6
Count of hyponyms	8.3	8	8.1
Count of syllables	6.7	6.6	6.6
Count of hypernyms	4.8	5	4.9
Length	4.1	4.1	4.1
Count of consonants	6.5	6.5	6.5
Count of vowels	3.8	3.5	3.6
Count of consonant conjuncts	4.8	4.6	4.7

Table 4.12: Aggregate of the feature importance values for every feature across all models for Approach 3

Feature	Importance Value based on		Aggregate
	Accuracy	Macro-F1	
Count of synonyms	5.1	5.8	5.4
Count of synsets	8.1	8.1	8.1
Frequency	10.6	10.6	10.6
Count of hyponyms	8.3	8	8.1
Count of syllables	6.6	6.6	6.6
Count of hypernyms	4.8	5	4.9
Length	4.1	4.1	4.1
Count of consonants	6.5	6.5	6.5
Count of vowels	3.8	3.5	3.6
Count of consonant conjuncts	4.8	4.6	4.7

During the process of determining the importance values, the researcher made the following observations:

- The accuracy of the model that was trained using the XGBoost classifier in Approach 1 was unaffected by the number of vowels.
- The macro-F1 value of the model that was trained using the XGBoost classifier in Approach 1 was not affected by the count of consonant conjuncts and the length of the word.

The researcher developed Method 2 in response to these observations, which show that the significance values of one feature vary drastically across different models. In Method 2, the researcher examined the accuracy, RUC scores, and macro-F1 scores of all the tree-based models. The results were contrasted with ALL 0 and ALL 1 as the baselines. Table 4.13 lists the metrics for Approach 1.

Table 4.13: Performance metrics for Approach 1

Classifier	Macro-F1	Accuracy
Baseline - ALL PREDICTIONS 0	0.22	0.81
Baseline - ALL PREDICTIONS 1	0.08	0.19
Nearest Centroid	0.26	0.47
Support Vector	0.26	0.47
Random Forest	0.28	0.65
Extra Trees	0.28	0.65
Gradient Boosting	0.27	0.59
XGB	0.27	0.59
Ada Boost	0.27	0.59
Decision Tree	0.28	0.66

As can be seen, no model has any discernible improvement over the baseline. It is evident that the accuracy and macro-F1 scores either fall below or remain relatively constant from the baseline. Due to the imbalance in the dataset produced by Approach 1, the researcher constructed a precision-recall curve, as seen in Figure 4.5. As shown in Figure 4.6, the researcher also compared it to the precision-recall curves for Approach 2.

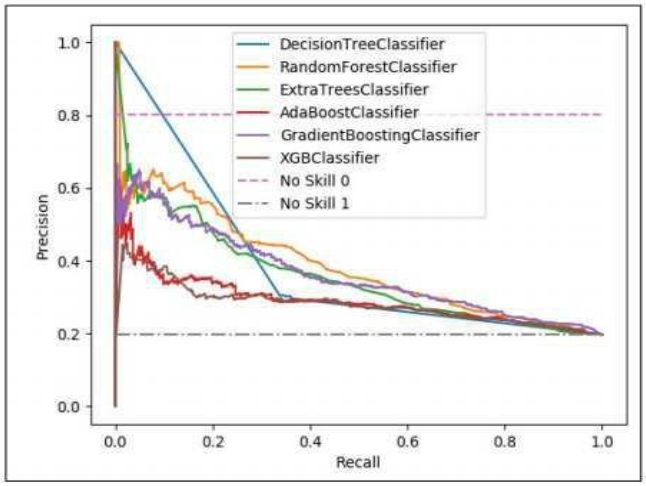

Figure 4.5: Precision-Recall curve for each tree-based model in Approach 1

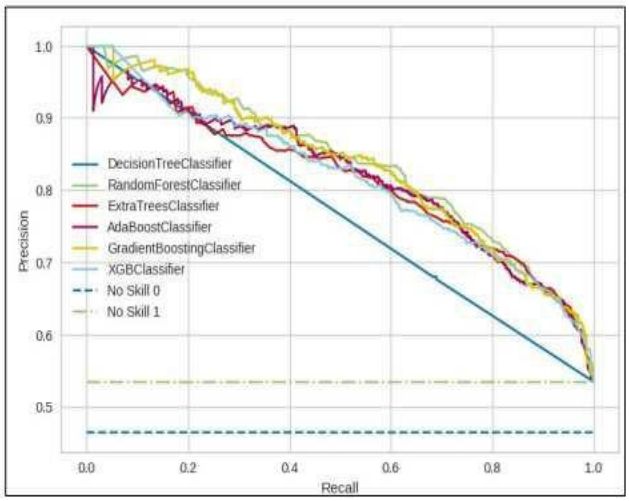

Figure 4.6: Precision-Recall curve for each tree-based model in Approach 2

Unlike Approach 2, there is no common point between precision and recall at which any model performs well, as shown by the plot for Approach 1. Because of this, the researcher chose to eliminate Approach 1 in the light of the outcomes of all the models. Despite the fact that there was no significant difference between these classifiers, the researcher found that the ensemble classifiers outperformed the decision tree.

Table 4.14 and Table 4.15 provide the metrics for Approach 2 and 3, respectively, and Figure 4.7 and Figure 4.8, respectively show the ROC curves for Approaches 2 and 3, respectively.

Table 4.14: Performance metrics for Approach 2

Classifier	Macro-F1	Accuracy
Baseline – ALL PREDICTIONS 0	0.16	0.47
Baseline - ALL PREDICTIONS 1	0.17	0.53
Nearest Centroid	0.32	0.63
Support Vector	0.29	0.62
Random Forest	0.36	0.73
Extra Trees	0.36	0.72
Gradient Boosting	0.37	0.75
XGB	0.37	0.73
Ada Boost	0.37	0.74
Decision Tree	0.33	0.67

Table 4.15: Performance metrics for Approach 3

Classifier	Macro-F1	Accuracy
Baseline - ALL PREDICTIONS 0	0.19	0.60
Baseline - ALL PREDICTIONS 1	0.14	0.40
Nearest Centroid	0.31	0.61
Support Vector	0.29	0.62
Random Forest	0.35	0.71
Extra Trees	0.34	0.70
Gradient Boosting	0.35	0.72
XGB	0.34	0.70
Ada Boost	0.35	0.72
Decision Tree	0.32	0.65

The values indicate that the results on Approach 2 are marginally better that the results on Approach 3. However, this observation does not serve as the basis to choose Approach 2 over Approach 3.

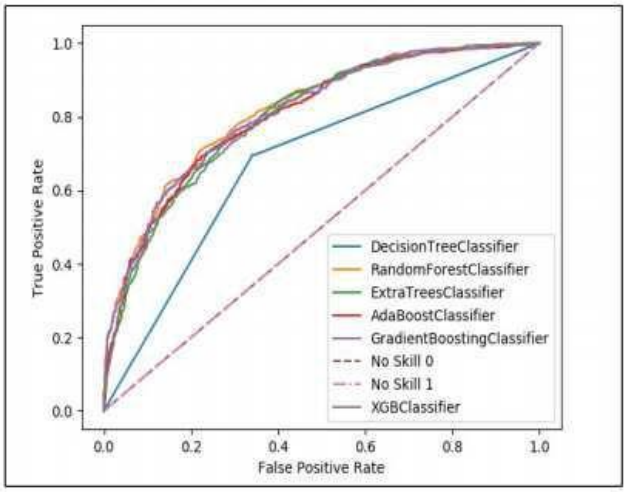

Figure 4.7: ROC curve for tree-based models in Approach 2

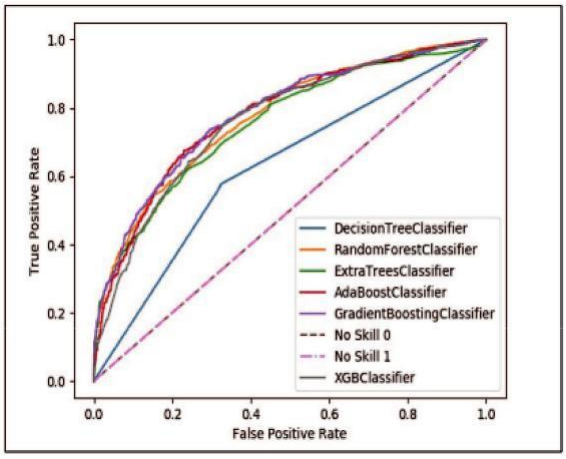

Figure 4.8: ROC curve for tree-based models in Approach 3

The probabilities of each class returned by all the models were then summed, and the researcher used a soft voting method on Approach 2 to choose which class should be retained. The equation associated with the soft voting method is as follows:

$\hat{y} = \text{argmax}_i \sum_{j=1}^{M} w_i p_{ij}$, where $\hat{y}$ is the predicted class, M is the count of models in the ensemble, w_j is the weight assigned to the prediction of the j^{th} model, and p_{ij} is the probability that is predicted by the j^{th} model for class i, where i is 0 or 1 (Mohammed & Kora, 2022).

The label that received the most votes among those created by ensemble classifiers and tuned ensemble classifiers was chosen. Although the researcher used randomised search to tune the models, the researcher found that there was no discernible difference between the output generated by the original model and that generated by the tuned models. Randomized search tuning is a hyperparameter tuning technique used to efficiently explore the hyperparameter space of a model. This method samples a fixed number of hyperparameter combinations randomly.

The label with the most votes among those created using the ensemble classifiers and the tuned ensemble classifiers was then chosen using a soft voting classifier. In a soft voting classifier, each participant classifier provides probability estimates for each class, and the final prediction is made based on the weighted average of the individual probabilities. This method considers the confidence or certainty of each classifier in their predictions in addition to the majority votes. As opposed to a hard voting classifier, wherein each classifier in the ensemble votes for a class and the class with the most votes is chosen as the final prediction, a weighted average method is used in soft voting, with each classifier contributing in proportion to its level of confidence. This enables more informative classifiers to influence the outcome more strongly.

Table 4.16 displays the results for the Area under the ROC Curve (AUC) scores for Approach 2.

Table 4.16: AUC Scores for Approach 2

Model	Approach 2
Ada Boost	0.80
Tuned Ada Boost	0.80
Extra Trees	0.80
Tuned Extra Trees	0.81
Gradient Boosting	0.81
Tuned Gradient Boosting	0.81
Random Forest	0.81
Tuned Random Forest	0.82
XGBoost	0.82
Tuned XGBoost	0.80
Soft Voting	0.82

Table 4.17 displays the results for the Area under the ROC Curve (AUC) scores for Approach 3.

Table 4.17: AUC Scores for Approach 3

Model	AUC Score
Ada Boost	0.78
Tuned Ada Boost	0.78
Extra Trees	0.76
Tuned Extra Trees	0.76
Gradient Boosting	0.78
Tuned Gradient Boosting	0.76
Random Forest	0.77
Tuned Random Forest	0.79
XGBoost	0.79
Tuned XGBoost	0.78
Soft Voting	0.79

The researcher found that the AUC score of the soft voting classifier was higher than the combined values of each classifier, for both, Approach 2 as well as Approach 3.

The next goal was to find the relevant features. The researcher estimated the significance values of the features in the models for each prediction using the classifiers that provided the accurate prediction. The ranks from the permutation feature importance approach were combined with the exhaustive feature set to produce the importance values. The same strategy that was used in Method 1 was applied to calculate the importance value. As can be seen in Table 4.18, the features were organised in decreasing order of importance value, for Approach 2 and Approach 3. Since no model could produce a result that was better than the baseline in Approach 1, the researcher eliminated Approach 1.

The features analysis provided convincing insights into the variables affecting word complexity in the context under study. Frequency was clearly the most important characteristic, coming in first place in the hierarchy of importance. Other features, such as length, synonyms, and consonant conjuncts, showed lower importance values. Interestingly, the count of vowels and the count of syllables were found to have converged in significance, with both factors receiving nearly identical ranks.

Table 4.18: Features ranked by relevance in decreasing order of their importance value

Feature	**Feature Importance Value**	**Feature Rank**
Frequency	1.15	I
Count of hyponyms	1.04	II
Count of syllables	1.036	III
Count of vowels	1.033	IV
Count of consonants	1.027	V
Count of synsets	1.023	VI
Count of hypernyms	1.022	VII
Length	1.017	VIII
Count of synonyms	1.013	IX
Count of consonant conjuncts	1.002	X

Conversely, features like the number of consonants, synsets, and hypernyms were ranked as relatively less relevant, indicating that they had less of an impact on word complexity. As can be seen, both approaches produced identical ranks. As a result, the researcher has solid proof that a word's frequency is a key predictor of its complexity. This is consistent with findings from research done on non-Indian languages.

4.3.2 Selection of Labeling Approach and Classifier Type

The researcher calculated the performance of six models - AdaBoost, Random Forest, Gradient Boosting, Extra Trees, XGBoost, and Decision Tree, on the datasets generated using two approaches. In Approach 1, the researcher labelled a word as complex if at-least two annotators assigned a rating of 3 or lower. In Approach 2, the researcher labelled a word as complex if the average rating received by it was 3 or lower. Since the dataset that was created using Approach 1 was not balanced, the researcher trained the models on an oversampled as well as an undersampled dataset. The researcher also considered the AUC score and the macro-F1 score because accuracy alone cannot be relied upon. The researcher chose to eliminate Approach 1 from the rest of the study because Approach 2 gave a better performance. Since Approach 2 and 3, gave similar performances in the previous experiment, the research chose Approach 2, as average gives an accurate measure of the central value from a list of values. Also, since the data, that is, the ratings were not categorical, it made sense to use the average of the ratings. The complexity, i.e., difficulty of a word in other languages has been shown to be highly correlated with its frequency (Bingel et al., 2016). The researcher obtained similar results as the outcome of the feature analysis experiment described in the previous sub-section. As a result, the researcher trained the models on datasets of various sizes in order to evaluate their performance. The researcher used two datasets to train the models: the dataset at hand and another that combined the AI4Bharat corpus with frequency values that were derived from the corpora at hand (Kunchukuttan et al., 2020). The AI4Bharat corpus is a project to compile extensive, general-domain corpora for Indian languages. It has 2.7 billion words for ten Indian languages spread over two language groups. Table 4.19 displays the outcomes of the performances for Approach 2.

Table 4.19: Percentage improvement in performance when the present corpus and the AI4Bharat corpus are combined to determine the frequency of words

Model	Approach 2		
	AUC Score	**Macro F1 Score**	**Accuracy**
AdaBoost	3.00	2.28	2.42
Decision Tree	2.13	2.54	2.71
Extra Trees	1.70	1.63	1.73
Gradient Boosting	1.74	1.06	0.67
Random Forest	2.04	2.27	2.02
XGBoost	1.81	0.15	1.34

As can be seen, the changes in performance values are positive. To examine how the removal of a specific feature will affect the performance of the models, the researcher conducted a feature ablation study. In a feature ablation study, the main objective is to systematically evaluate the contribution of each feature or group of features to the model's performance. The researcher measured accuracy, macro-F1, and AUC score for each model using Approach 2, and then eliminated one feature at a time. Since ablation study was used multiple times in this study, the researcher made the code open source under the GNU GPL v3 open source license. The code for the study is available for use and modification on GitHub[1].

Table 4.20 showcases the results of the ablation study on AdaBoost and Extra Trees models, and Table 4.21 showcases the results of the ablation study on the rest of the models.

[1]https://github.com/gayatrivenugopal/ablation-study

Table 4.20: Ablation study results for Approach 2 for AdaBoost and Extra Trees

Model	Features	Accuracy	Macro-F1	AUC
AdaBoost	All	0.72	0.71	0.78
	All-length	0.72	0.71	0.77
	All-Count of synsets	0.72	0.70	0.78
	All-Count of synonyms	0.71	0.70	0.78
	All-Count of consonants	0.72	0.70	0.77
	All-Count of vowels	0.72	0.70	0.77
	All-Count of hypernyms	0.71	0.70	0.77
	All-Count of hyponyms	0.71	0.70	0.77
	All-Count of consonantconjuncts	0.70	0.68	0.77
	All-Count of syllables	0.72	0.70	0.78
	All-frequency	0.67	0.63	0.70
Extra Trees	All	0.70	0.67	0.75
	All-length	0.70	0.68	0.74
	All-Count of synsets	0.70	0.68	0.75
	All-Count of synonyms	0.70	0.68	0.75
	All-Count of consonants	0.70	0.67	0.75
	All-Count of vowels	0.71	0.69	0.75
	All-Count of hypernyms	0.70	0.67	0.75
	All-Count of hyponyms	0.70	0.68	0.75
	All-Count of consonantconjuncts	0.70	0.68	0.75
	All-Count of syllables	0.69	0.67	0.74
	All-frequency	0.67	0.64	0.69

Table 4.21: Ablation study results for Approach 2

Model	Features	Accuracy	Macro-F1	AUC
Gradient Boosting	All	0.72	0.71	0.77
	All-length	0.72	0.70	0.78
	All-Count of synsets	0.72	0.71	0.77
	All-Count of synonyms	0.72	0.70	0.77
	All-Count of consonants	0.72	0.70	0.78
	All-Count of vowels	0.72	0.70	0.77

Model	Features	Accuracy	Macro-F1	AUC
	All-Count of hypernyms	0.71	0.70	0.77
	All-Count of hyponyms	0.71	0.70	0.77
	All-Count of consonantconjuncts	0.71	0.70	0.77
	All-Count of syllables	0.72	0.70	0.77
	All-frequency	0.72	0.70	0.77
XGBoost	All	0.70	0.68	0.76
	All-length	0.70	0.68	0.76
	All-Count of synsets	0.70	0.68	0.76
	All-Count of synonyms	0.70	0.68	0.76
	All-Count of consonants	0.70	0.68	0.76
	All-Count of vowels	0.70	0.68	0.76
	All-Count of hypernyms	0.70	0.68	0.76
	All-Count of hyponyms	0.70	0.69	0.75
	All-Count of consonantconjuncts	0.70	0.68	0.75
	All-Count of syllables	0.70	0.68	0.75
	All-frequency	0.66	0.61	0.69
Random Forest	All	0.689	0.693	0.77
	All-length	0.771	0.679	0.76
	All-Count of synsets	0.707	0.687	0.77
	All-Count of synonyms	0.695	0.676	0.76
	All-Count of consonants	0.701	0.680	0.76
	All-Count of vowels	0.701	0.682	0.77
	All-Count of hypernyms	0.706	0.687	0.76
	All-Count of hyponyms	0.706	0.686	0.76
	All-Count of consonantconjuncts	0.710	0.690	0.76
	All-Count of syllables	0.706	0.686	0.76
	All-frequency	0.673	0.644	0.71
Decision Tree	All	0.648	0.635	0.64
	All-length	0.644	0.631	0.64
	All-Count of synsets	0.651	0.637	0.64
	All-Count of synonyms	0.635	0.621	0.62
	All-Count of consonants	0.647	0.634	0.64

Model	Features	Accuracy	Macro-F1	AUC
	All-Count of vowels	0.648	0.635	0.64
	All-Count of hypernyms	0.645	0.631	0.64
	All-Count of hyponyms	0.639	0.624	0.63
	All-Count of consonantconjuncts	0.627	0.614	0.62
	All-Count of syllables	0.633	0.620	0.62
	All-frequency	0.63	0.61	0.62

The researcher found that the majority of the models' performances varied more when frequency was removed. But this was not true for all the models. The researcher investigated the effectiveness of the models by keeping frequency as the only feature in the dataset because the other features do not make a significantly meaningful contribution. However, the researcher discovered that the models' performance dropped, thus proving that frequency alone is insufficient as a predictor. The outcomes are displayed in Table 4.22. The researcher's attention was drawn to the AUC score among these metrics. The researcher discarded this option after discovering that the scores differed significantly when frequency was the only characteristic used.

The researcher also checked if the dataset was suitable for linear regression. However, the conditions for linear regression were not met for the available data. The labels and data points did not have a linear or a non-linear relationship. Figure 4.9 displays the correlation heatmap between the features and the label.

Table 4.22: Comparison of the performance of models built on datasets created using Approach 2 with all features vs models constructed on datasets with frequency

Model	Features	Accuracy	Macro F1 Score	AUC Score
AdaBoost	all	0.72	0.71	0.78
	frequency	0.69	0.33	0.74
Extra Trees	all	0.70	0.67	0.75
	frequency	0.66	0.33	0.66
Gradient Boosting	all	0.72	0.71	0.78
	frequency	0.70	0.34	0.74
XGBoost	all	0.70	0.68	0.75
	frequency	0.69	0.34	0.74
Random Forest	all	0.69	0.70	0.77
	frequency	0.66	0.33	0.68
Decision Tree	all	0.65	0.64	0.64
	frequency	0.66	0.33	0.64

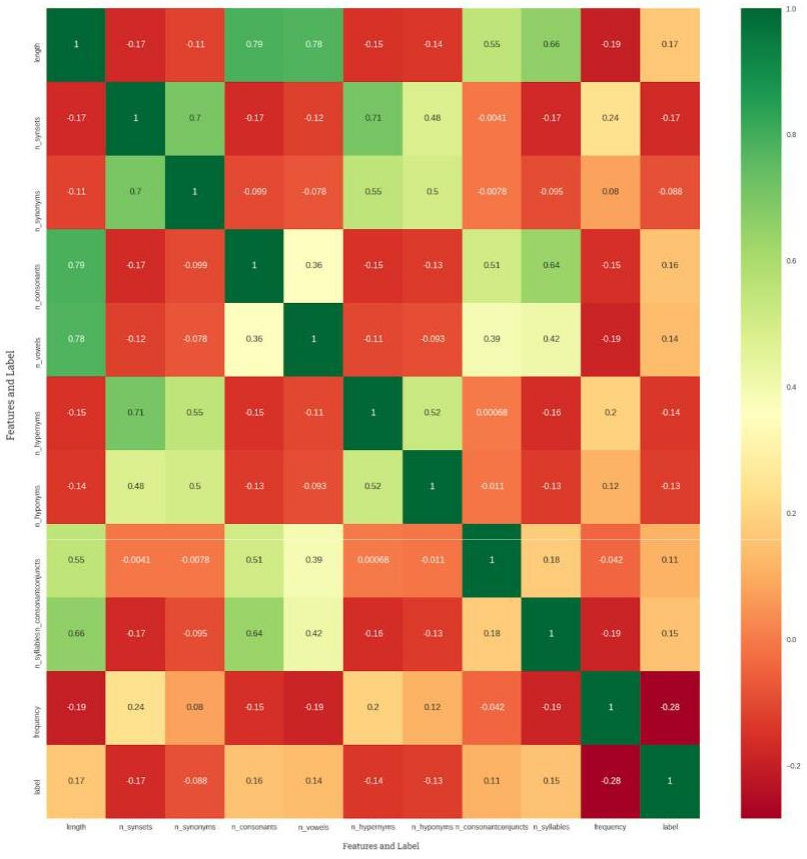

Figure 4.9: Heatmap depicting the correlation between the features and the label

As can be seen, there is no significant relationship between any one individual feature and the label, owing to which this problem cannot be solved using a traditional programming approach. This justifies the machine learning approach that is used for this research.

In order to identify the ideal hyperparameters, the researcher used the dataset created by Approach 2 and implemented grid search. Considering that the base classifiers performed similarly, the researcher developed a voting classifier. Decision trees were not used in the creation of the voting classifier since ensemble models gave a better performance. As the researcher could not run grid search on Gradient Boosting and Extra Trees models using AUC scores, these models were not taken into account in this scenario. Table 4.23 displays the AUC values for the non-tuned models, that is, models created using default parameters, and the models tuned using the grid search hyperparameter tuning method.

Table 4.23: AUC scores for non-tuned and tuned models

Model	**AUC Score**
Random Forest	0.78
AdaBoost	0.77
XGBoost	0.78
Tuned Random Forest	0.78
Tuned AdaBoost	0.79
Tuned XGBoost	0.77
Soft Voting Classifier	0.79

The researcher created a soft voting classifier rather than a hard voting classifier because the former takes into account the accuracy of each classifier's prediction.

4.3.3 Word Embeddings as Features

In order to evaluate the use of word embeddings as features, the researcher compared the outcomes of the models employing word embeddings, frequency, and lexical variables, using different combinations of features. In Table 4.24, Table 4.25, Table 4.26, and Table 4.27, it can be seen how the models performed with the various dataset types. Although AUC score and F1

score were the primary metrics of concern, the researcher also included the outcomes of additional performance measures in the report.

Table 4.24: Results of the soft voting classifier using a dataset with only frequency and lexical features

Metric	Train	Test
AUC	0.774	0.727
Macro F1	0.344	0.325
F1	0.601	0.555
Precision	0.660	0.624
Recall	0.551	0.50
Accuracy	0.707	0.667

Table 4.25: Results of the soft voting classifier using a dataset with solely pre-trained embeddings

Metric	Train	Test
AUC	0.773	0.733
Macro F1	0.341	0.315
F1	0.558	0.469
Precision	0.694	0.641
Recall	0.467	0.369
Accuracy	0.704	0.652

Table 4.26: Results of the soft voting classifier using a dataset with pre-trained embeddings and word frequency

Metric	Train	Test
AUC	0.793	0.759
Macro F1	0.349	0.336
F1	0.607	0.563
Precision	0.682	0.675
Recall	0.547	0.484
Accuracy	0.717	0.689

Table 4.27: The effectiveness of the soft voting classifier on a dataset with lexical characteristics, word frequency, and pre-trained embeddings

Metric	Train	Test
AUC	0.803	0.766
Macro F1	0.361	0.340
F1	0.642	0.590
Precision	0.706	0.672
Recall	0.590	0.525
Accuracy	0.737	0.696

A comparison of the effectiveness of the models trained on the various datasets is shown in Figure 4.10. When the researcher combined frequency, lexical, and pre-trained embedding characteristics, the researcher found that the classifier performed at its best. This suggests that the combination of these three different feature types improves the model's capability to identify and categorise complex words. The performance of the models trained on pre-trained word embeddings and frequency, as well as the models trained on pre-trained word embeddings, frequency, and lexical features, respectively, both outperform the models trained on only lexical features and only word embeddings. In the context of the classification challenge, the complementing nature of frequency, lexical, and pre-trained embedding features is indicated by this hierarchy in performance. The F1 score of the model trained on all features was greater than the model trained on pre-trained word embeddings and frequency, despite the fact that the AUC score and Macro-F1 score of these two models were comparable. This implies that the addition of lexical information improves the model's accuracy and recall in the particular context of classification, resulting in a better F1 score.

In conclusion, integrating frequency, lexical, and pre-trained embedding characteristics led to the classifier's best performance. The comparative study of various feature combinations emphasises how crucial it is to take into account a variety of feature kinds for reliable model performance. The ablation test explores lexical elements in greater detail and offers a more sophisticated perspective of their influence on the identification of complex words.

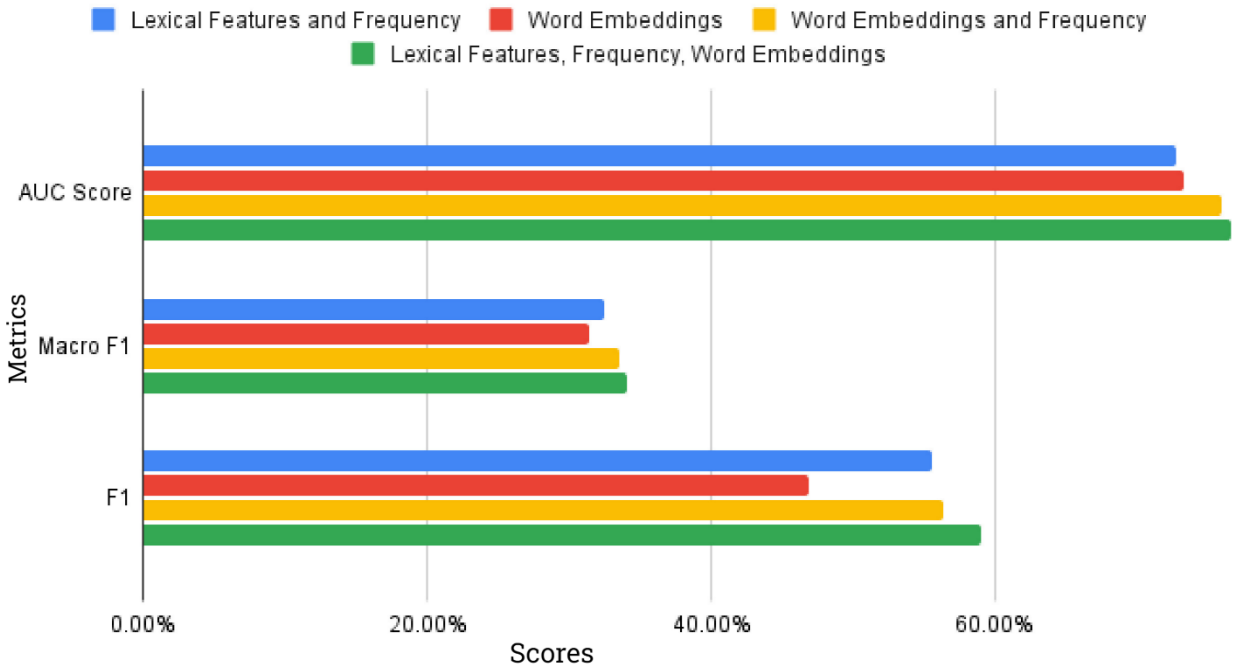

Figure 4.10: Performance evaluation of models trained on datasets with different features

The outcomes of the ablation test performed on the classifier trained on all lexical features are shown in Table 4.28. Although the researcher saw a significant improvement when frequency was added back to the dataset, the performance improved when the other features were added. This showed that while one feature by itself may not make a significant contribution, when combined with the other features, it can help identify complex words.

For the concluding experiment, the researcher built a neural network with all the features, including three fully connected layers, ReLU activation functions at each layer, and a binary cross entropy loss function that was trained over 50 iterations. Note that this experiment was not included in the scope of the study. However, the researcher intended to test neural networks at a basic level to solve this problem. The researcher observed that the average loss at the end of every epochs was very high (0.69), and that the accuracy was lower than that achieved using the ensemble soft voting classifier (0.58). Therefore based on these results, the researcher concludes that the ensemble soft voting classifier would be a better predictor of complex words, as compared to neural networks.

Table 4.28: Ablation test results

Dataset	AUC Score	Macro F1	F1
All	0.77	0.34	0.59
All-len	0.76	0.35	0.61
All-synsets	0.77	0.35	0.60
All-synonyms	0.76	0.34	0.58
All-consonants	0.77	0.34	0.59
All-vowels	0.76	0.34	0.58
All-hypernyms	0.76	0.34	0.58
All-hyponyms	0.76	0.34	0.59
All-consonantconjuncts	0.77	0.34	0.59
All-syllables	0.77	0.34	0.59
All-frequency	0.74	0.32	0.32

4.3.4 End-to-End Lexical Simplification Pipeline

This section describes the end-to-end lexical simplification pipeline for providing simpler alternatives to complex words in a Hindi sentence.

The pipeline begins with a Hindi sentence as the input. For example, consider the following sentence as input:

पंडित जी ने कहा कि ब्रह्माण्ड में जिस मानव के पास सभी सुख सुविधाएं हैं, वह भी सुखी नहीं है।

English translation: Pandit ji said that even a human being who has all the comforts and facilities in the universe is not happy.

The first stage of the pipeline is pre-processing the sentence to remove irrelevant characters. Here, ‘,’ and ‘।‘ will be removed as part of this process. The next step is to split the sentence at space and create a list of words (or tokens), that is, tokenisation. The set of steps is shown in Figure 4.11

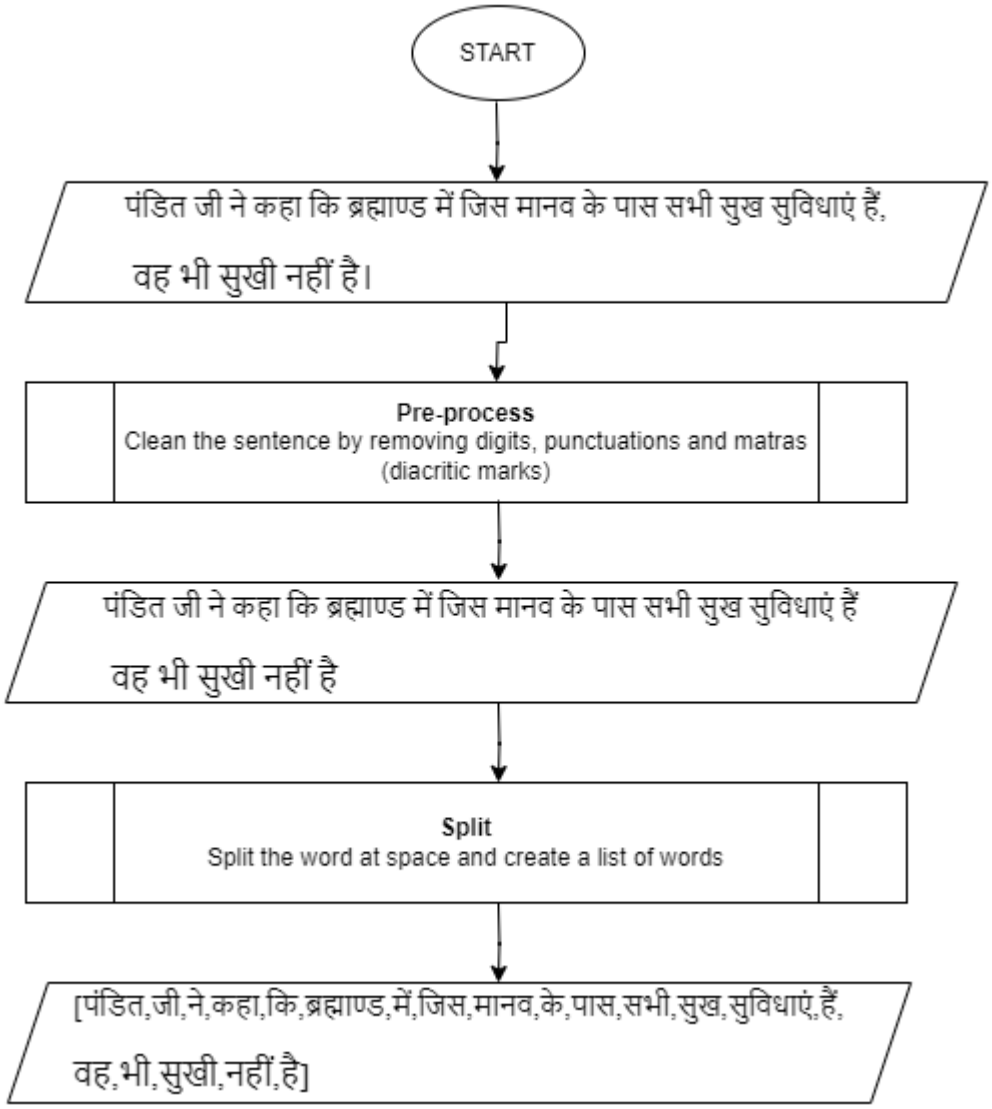

Figure 4.11: Pre-processing phase of the lexical simplification pipeline

The next stage of the pipeline is to fetch each word from the list and process it by fetching its root form, that is, lemmatisation, ensuring that a stop lemma is not being processed, and then fetching the synsets of the word from the Hindi WordNet (Panjwani et al., 2018). The steps are shown in Figure 4.12.

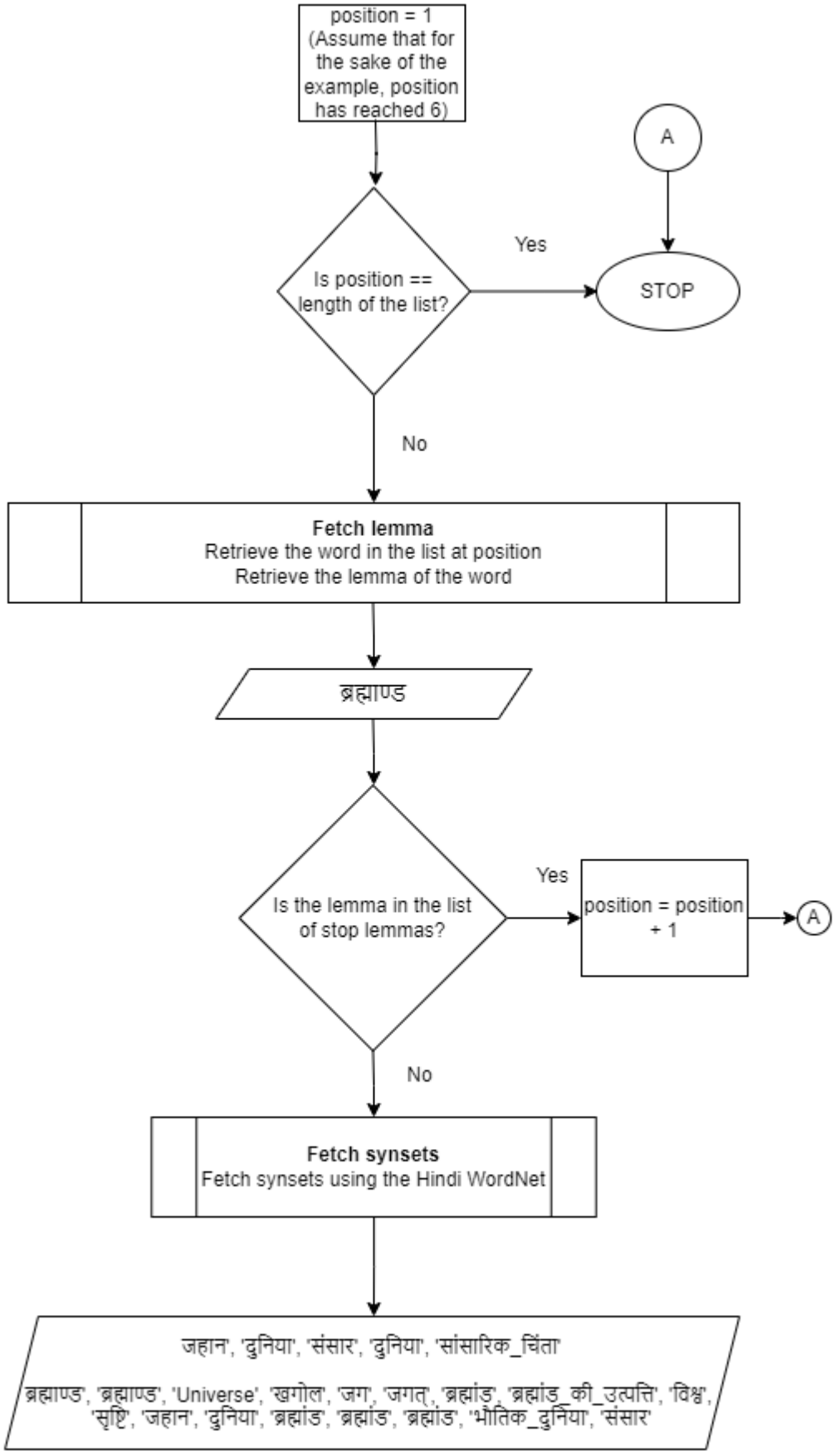

Figure 4.12: The lemmatisation and synset fetching stage in the lexical simplification pipeline

The next step of the pipeline is to prepare the word in the form of a record in the dataset that was used to build the classifier. For this, the system will vectorize the word, create a group for the word, consisting of the word itself, along with all its synonyms, and fetch the lexical feature values for all the word and its synonyms. Then the system applies synset based normalisation method to all the values and updates the values of the lexical features of the word. These values, along with the embeddings of the word form a record, which is then added to data structure for further processing. This data structure is then consumed by prediction stage, wherein the system loads the complex word classifier and processes the data structure to predict a binary complexity of the word, along with it's probability of prediction. The steps are shown in Figure 4.12.

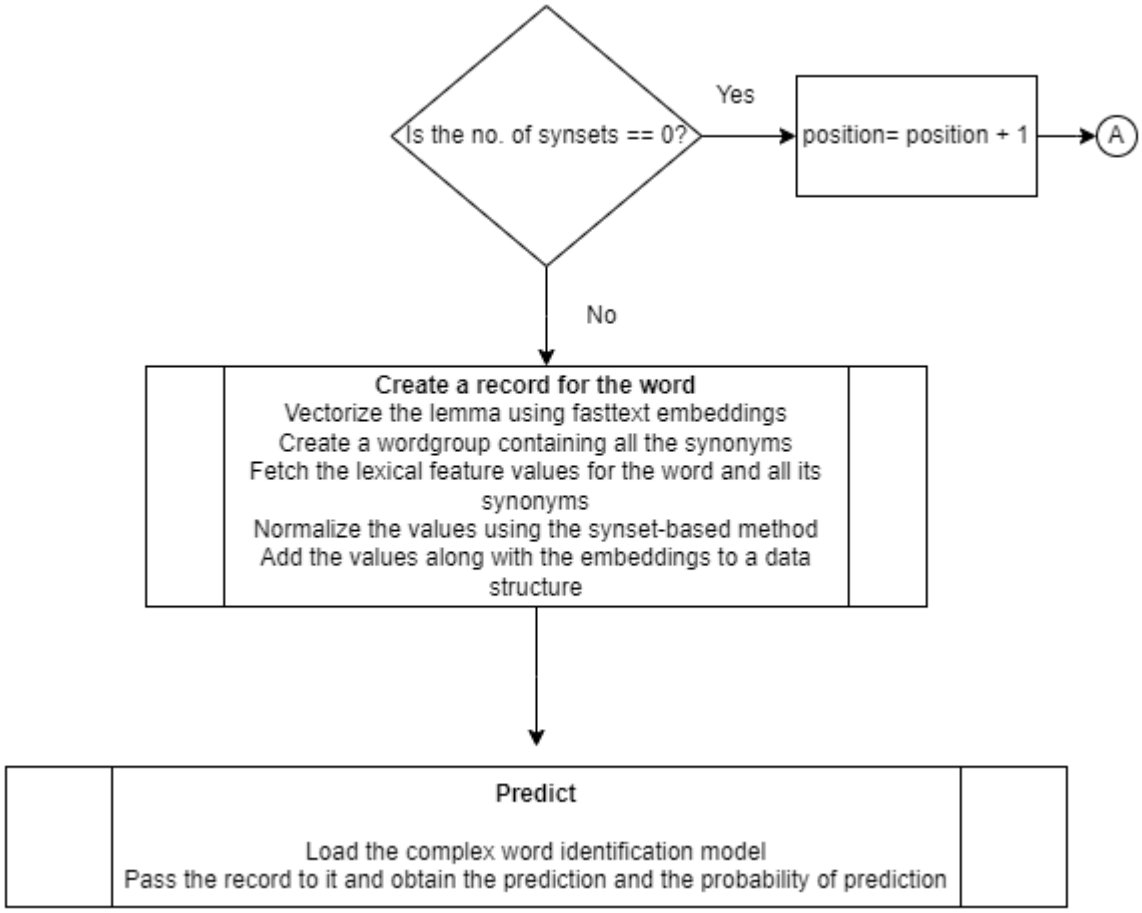

Figure 4.12: The preparation of the word to be consumed by the prediction process, and the prediction process in the lexical simplification pipeline

After complex word identification, the next task is to disambiguate the word. If the system predicts the word to be complex, and if the probability of this prediction is greater than or equal to 0.5, the next step is activated. In the word sense disambiguation phase, the system fetches the synset embeddings for the target word using ARES embeddings (Scarlini et al., 2020b). Then it calculates the cosine similarity of the synset embeddings and the target word embeddings that are retrieved from the mBERT model. The system then finds the synset with the highest similarity, and fetches the senses for the synset using the synset ID from BabelNet.

An example of the selected synset is as follows:

ब्रह्माण्ड', 'ब्रह्माण्ड', 'Universe', 'खगोल', 'जग', 'जगत्', 'ब्रह्मांड', 'ब्रह्मांड_की_उत्पत्ति', 'विश्व', 'सृष्टि', 'जहान', 'दुनिया', 'ब्रह्मांड', 'ब्रह्मांड', 'ब्रह्मांड', 'भौतिक_दुनिया', 'संसार'

The final phase is lexical simplification, that is, providing simpler alternatives. Here, the system compares the probability of prediction that it obtained for the target word, with the probability of prediction of obtaining 1 as the prediction for each synonym. If the prediction for the synonym is 0, that the probability of obtaining 1 is less than that for the target word, then this synonym, along with the new minimum complexity probability is stored. This process is repeated for all the synonyms. In the end, the system would store the synonym for which the prediction was 0, and that has the lowest probability of obtaining 1 as the prediction, among all the other candidate synonyms. The pipeline then displays this synonym as the final output to the user.

A graphical representation of the last stage of the pipeline can be seen in Figure 4.13.

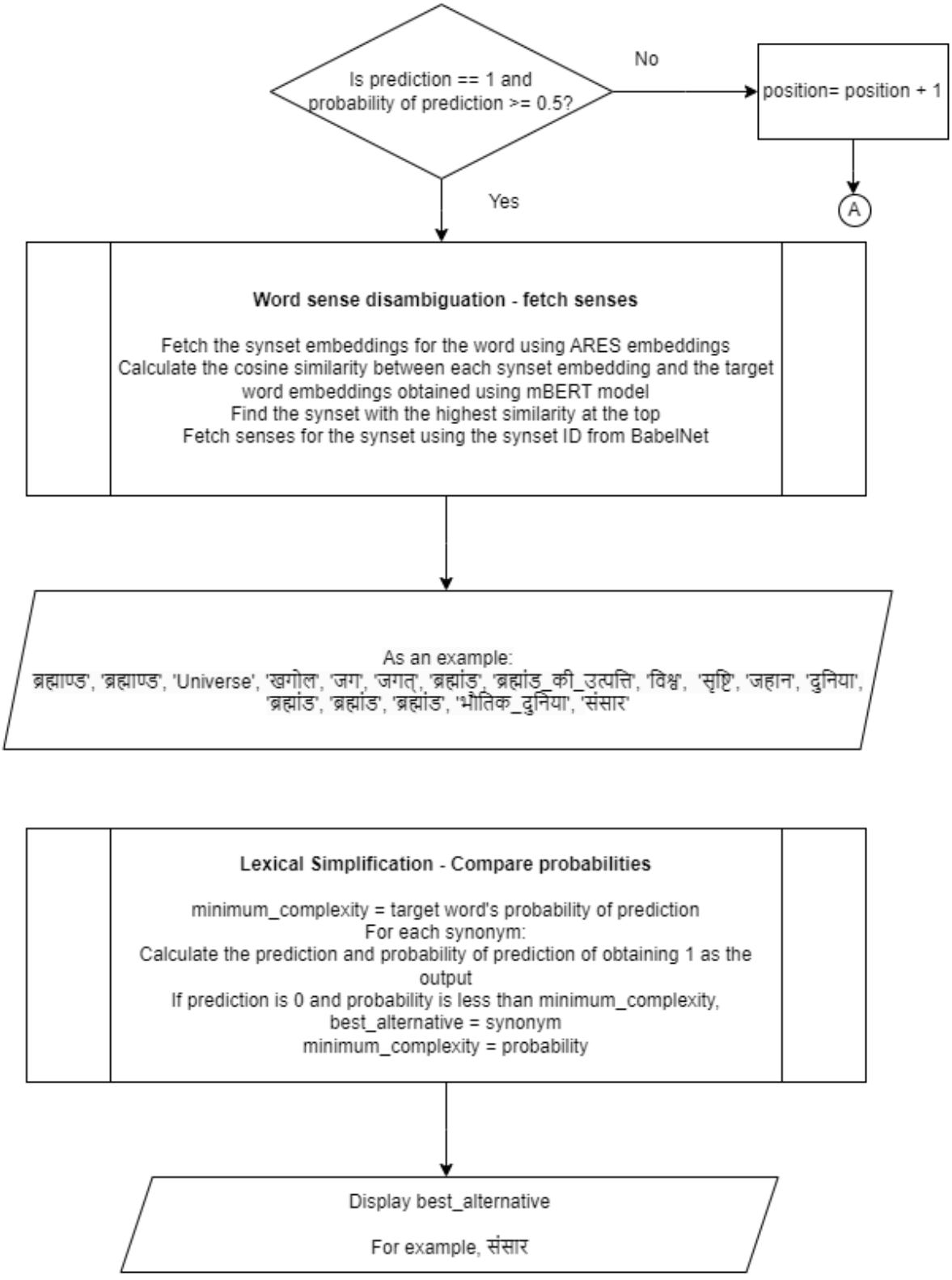

Figure 4.13: The final stage of the lexical simplification pipeline

Chapter 5

Conclusion

5.1 Concluding Remarks

The researcher believes that the culmination of this rigorous research journey in the field of Hindi complex word identification and lexical simplification marks a significant milestone in the areas of language technology, linguistic analysis, and text accessibility. The researcher's work helped not only understand and highlight the complexities inherent in the Hindi language but also charted a path toward enhancing communication efficiency and inclusivity for diverse audiences. In this chapter, the researcher reflects on the insights, implications, and future directions of Hindi lexical simplification. The investigation into the identification of Hindi complex words has revealed the intricate web of word complexity within the diverse linguistic environment of Hindi. As part of the research, a sizable dataset that was carefully assembled and curated to highlight the varied character of word complexity was created and annotated. A framework for detecting complex words has been produced as a result of combining traditional linguistic research with machine learning techniques. This framework will serve as the basis for future developments. The implications of the contributions extend beyond the confines of academic discourse, resonating with real-world applications that stand to transform how information is disseminated and understood within the Hindi-speaking population. The educational landscape, for instance, can leverage the research findings to craft learning materials that cater to diverse proficiency levels. By offering simplified versions of complex texts while retaining the core essence, educators can bridge the gap between introductory and advanced materials, fostering a smoother transition for learners. Moreover, the healthcare and legal sectors, which often grapple with intricate terminologies, can harness the methodologies used in this research to empower individuals with accessible information. Patients can make informed decisions about their health, while individuals navigating legal matters can grasp the intricacies of their rights and responsibilities.

The idea for this research originated from the need to improve the accessibility of text written in Hindi. The researcher set out to solve this problem by conducting a study to simplify words in a given Hindi sentence. The researcher studied existing literature and observed that there was

a dearth of resources in Hindi for text simplification, specifically, lexical simplification. The researcher attempted to replicate existing studies, but given the absence of comparable resources in Hindi, such as simple word lists, psycholinguistic databases, or the Hindi Wikipedia, the researcher decided against trying to replicate the methods. Therefore, the researcher aimed to create a corpus of Hindi text. The researcher created a corpus of aesthetic literature in order to train the model on generic sentences, used in Hindi literature, as opposed to texts in domains such as law, medicine, technology, etc. In order to compile a list of Hindi stop words, the researcher also aimed to analyse the stop words that were present in the corpus that the researcher created, as well as those found in publicly accessible lists and other corpora. This list could also be of use in studies beyond text simplification in Hindi.

The researcher used literature from both the fictional and non-fictional categories to accomplish the first goal. Using content from multiple sources, an aesthetics corpus of about 1,000 articles and 1,45,508 words was produced. The researcher was unable to find a large body of work by female authors in digital format that was available to the general audience. The corpus includes both, recent literature and stories that are more than a century old. Seven distinct corpora of texts were combined to form an extensive stop word list. The researcher also utilised the publicly accessible stop word lists. Stop word lists were created based on the word count in each corpus. It was discovered that there were discrepancies among all of the lists, even in the top ten stop words.

Through this research, the researcher advocates the usage of stop lemma lists rather than simple stop word lists in natural language processing tasks. Stop lemmas were found to be more reliable than stop words across several corpora, hence stop lemmas were taken into account in order to create a robust list. This was determined by word cloud density made using the top ten terms and top ten lemmas from the sources the researcher included. At the time the study was being conducted, such a list was not accessible to the general public. Only the texts that were compiled from the sources are included in this list. The researcher assessed the list's quality by lemmatizing, translating, and comparing the lemmas with the English stop words that are present in the NLTK Python package that is often used in English natural language processing tasks. Because it was necessary to be familiar with the language to determine the quality of the translation, the researcher chose English. The researcher discovered that nearly every lemma in the Hindi list has an equivalent in the English list. The list may be evaluated by using it to perform different natural language processing tasks and comparing the results of one particular

natural language processing task with relation to the use of stop words and stop lemmas in various languages as part of future research that stems from this study. Public access to the corpus, the metadata, and the analysis is accessible at https://github.com/gayatrivenugopal/hindi-corpus-stoplemmas.

The researcher devised a user study and determined the lexical features from which the researcher can extract values. Sentences from the Hindi text corpus were presented to the annotators after the corpus had been compiled and developed. Annotators were asked to annotate complex words and assign a ranking of simplicity to each synonym, as well as the complex word itself after being presented with its synonyms depending on the word they had chosen as complex in the sentence. The researcher created a dataset for Hindi complex word identification that is independent of context. Stop words were not included in this list of words. The 7,321 words in the dataset were annotated by 100 different annotators. The researcher's objectives were to develop the dataset and comprehend how various people perceive complex words. This study shows that in an Indian context with several regional languages, a person's native language may not be their preferred language. With 93.05% choosing English as their preferred reading language, 72% of annotators felt more at ease reading literature in a language other than their mother tongue. Notably, Hindi was not preferred by native Hindi speakers. 24% annotators favoured their region's official language, whereas 66% favoured English. This suggests that a person's preference for a language might be influenced by the area in which they have lived for the longest. In addition to years of academic education in a particular language, the most comfortable language may also depend on the language that is spoken in the environment where they resided for the most of their time. Overall, non-native speakers annotated more sentences than native speakers, but on average, native speakers annotated more sentences.

The number of terms annotated did not correlate with years of academic experience in Hindi, and there was little consensus among the annotators. Less words were marked by annotators who regularly read Hindi-language content, indicating that reading habits rather than formal education may be a more reliable measure of language proficiency. The researcher discovered a considerable difference between the feature values of a simple word and a complex word, proving the need for additional investigation into the function of these traits. Less than two-thirds of the 8,744 words were classified as simple in their lemmatized forms but as difficult in their non-lemmatized forms. This indicates that the lemmatized form had little effect on the perception of word complexity.

Using datasets created using two labelling techniques, the researcher assessed six models: AdaBoost, random forest, gradient boosting, extra trees, XGBoost, and decision tree. Whereas Approach 2 employed the average rating, Approach 1 classified a word as complex if at least two annotators gave it a score of three or lower. Approach 2 was preferred over Approach 1 because of its superior results. AUC and macro-F1 scores were used to assess the models. A model that is based on normalised features in relation to the target word and its synonyms was created. The model's predictions were derived using tree-based ensemble models. Although there is not a significant bias in favour of any particular user type, the model analysis reveals that there are significant variances between various categories. The results of the study showed that groups of users who were more accustomed to using the language in everyday situations had slightly higher agreement scores. In light of this, test sets created by merging the annotations of the annotators performed marginally better than those created by their counterparts in each area. The dataset may be used to construct a lexical simplification system that proposes simpler words to use in place of a difficult word and is publicly available at https://zenodo.org/record/5229160.

The researcher set out to find out if it is possible to determine how complex a word is in a given Hindi sentence using the criteria of conventional readability formulae, which are used to grade the readability of both English and non-English language material. Using methods such as the permutation feature importance, and the exhaustive feature selection method, the researcher determined the importance of features using metrics such as accuracy and macro-F1 scores. The researcher learned from the experiments that frequency affects both, the readability and the complexity of a word. The majority of readability formulae have concentrated on length; however, it does not rank highly among the important parameters for identifying complex words. The researcher found that the complexity of a word is not significantly affected by the count of consonant conjuncts, a distinctive characteristic of Hindi words. The number of syllables is a crucial component. The researcher found that an essential factor in determining a word's complexity, is the count of hyponyms it contains, that is, the count of words that fall under the same category as the target word. This was a parameter that was missing from the word lists and readability formulae. The findings make it clear that a readability metric for Hindi at the word level and a sentence level is required.

The results validate the possibility of creating efficient classifiers based on these features that can differentiate between simple and complex words. The findings support the applicability of these features in other linguistic contexts and are consistent with earlier studies on non-Indian languages. Using a variety of methods, the researcher identified the relevant features from eight different models, five of which were ensemble models based on trees. Only the ensemble models were considered for creating a complex word classifier as they outperformed the others. Gradient boosting and extra trees classifiers could not be tuned using grid search and the AUC metric. Therefore the remaining three models were tuned, and the tuned and non-tuned models were used to create a soft voting classifier. A soft voting classifier was employed because it outperformed all other models in terms of AUC scores. It was a combination of Random Forest, AdaBoost, XGBoost, Tuned Random Forest, Tuned AdaBoost, and Tuned XGBoost, where models were tuned using grid search. It produced an AUC score of 0.79 on the training set and 0.77 on the test set. After including word embeddings, frequency and lexical features, the AUC score of the soft voting classifier was 0.80 on the training set and 0.76 on the test set. The researcher was able to discern complex words from their simpler equivalents with good accuracy of 0.70 on the test set by combining linguistic features and machine learning models. The researcher built a neural network with all the features, including three fully connected layers, ReLU activation functions at each layer, and a binary cross entropy loss function that was trained over 50 iterations. The proposed ensemble classifier outperformed the neural network, achieving an accuracy of 0.58. In contrast, the neural network had a high average loss of 0.69 and lower accuracy after 50 iterations. These results substantiate the superiority of the ensemble classifier in predicting complex words. This accomplishment highlights the potential of combining linguistic knowledge with computational methods to create a hybrid methodology that is in line with the changing field of natural language processing.

5.2 Contributions and Limitations

The key contributions of this doctoral research are as follows:

- Stop lemma list consisting of 311 stop lemmas in Hindi

  The stop lemma list is a one of its kind unique list, that is, it contains the root forms of words. The lemma list was formed by compiling multiple resources, including existing

stop words lists, and by extract stop words and subsequently their lemmas, from various corpora. It has been evaluated by comparing the list with the list of English stop words provided by the NLTK Python package. This list has been released under an open source license, and is available to the public at https://github.com/gayatrivenugopal/hindi-corpus-stoplemmas.

- Aesthetics corpus

  The researcher created a corpus containing 145,508 unique words 118,266 unique lemmas. The corpus spans a period of 100 years and contains Hindi sentences from short stories, novels, biographies, and poems. The corpus can be used for conducting various types of studies on Hindi text. The corpus is available for public use at https://github.com/gayatrivenugopal/hindi-corpus-stoplemmas/tree/master/aesthetics%20corpus.

- CWID-hi dataset

  The dataset that was also tested for bias contains 7,321 words and was annotated by 100 native as well as non-native Hindi speakers. It can be used to extend the work on Hindi lexical simplification, for instance, to create a personalized simplification system. The dataset is available at https://zenodo.org/record/5229160.

- Synset-based normalization of features

  The researcher normalized features in the dataset by considering the features of the target word and that of the synonyms in its synset, as opposed to normalizing the entire dataset, which would include the values of unrelated words as well. Therefore, the researcher avoided comparing the value of the target word with that of words that were irrelevant with respect to the target word.

- Word Sense Disambiguation in Hindi using ARES Sense Embeddings

  For determining the sense of a target word in a given Hindi text, this study used ARES-multilingual embeddings for fetching sense embeddings together with a variety of pre-trained embeddings. The researcher presented a unique approach for word sense disambiguation that had not been explored prior to this study.

- Soft voting classifier for complex word identification and lexical simplification

  The researcher created a classifier that combined the outputs of several tree-based ensemble models and obtained satisfactory results. The results may not seem very good as compared to typical performances of machine learning models. However, the researcher must consider the fact that the researcher is trying to make a machine predict the complexity of a word, which is extremely subjective to the reader. Considering this, and the results of the systems created till date, this system is at par, and in certain cases better than other systems that were created in resource-constrained environment as ours.

This study on lexical simplification is the first of its kind in Hindi. The researcher's approach reflects a departure from mere word substitution and delves into the realm of nuanced meaning preservation. This nuanced approach has far-reaching implications, particularly in domains where precision and clarity are paramount, such as education, healthcare, and legal documentation.

Owing to this, it has its own set of limitations. The researcher was not able to get very good results because a word's complexity is highly dependent on the reader's vocabulary. It is challenging to gauge a word's complexity due to the linguistic range of the participants and potential readers. However, in an effort to promote the use of plain language, the researcher was determined to contribute to the field of Hindi text simplification.

The prospects for further research and innovation in the areas of Hindi complex word identification and lexical simplification come into prominence as the researcher look to the future. An intriguing new approach that enables academics to compare Hindi and other languages to find common patterns of word complexity and simplification is the integration of cross-lingual resources. Collaboration between linguists, academicians, and technology can open the door to dynamic simplification models that adjust to various audiences, promoting inclusion and accessibility on a large scale. A more thorough method of simplification that can accommodate numerous language variations and geographical variances within Hindi is also possible by exploring dynamic simplification, which may be adjusted to varied sociocultural circumstances.

5.3 Future Scope

As improvements in natural language processing continue to transform communication, education, and technology, the future potential of complex word identification is enormous. Natural language processing models are anticipated to play a key role in improving language accessibility and understanding as they develop in sophistication and refinement. Numerous fields may benefit from the use of complex word identification. Intelligent tutoring systems in the field of education could make use of this technology to adapt the learning materials to the comprehension capacities of specific students, resulting in efficient and individualised learning experiences. A wider audience, including people with different language backgrounds or cognitive skills, may find online information easier to comprehend as a result of the automation of complex word identification. Communication gaps and inclusion could be greatly improved by integrating complex word identification into digital interfaces, language learning platforms, and assistive technology. Furthermore, the future of complex word identification shows potential for fostering innovation in sectors that depend on clear and concise communication. Natural language processing-powered technologies may help in the creation of documents in the fields of business writing, corporate communications, and legal writing that convey complex ideas while being simple and clear. By bridging the gap between specialists and non-experts, this may improve effective communication inside organisations. Furthermore, as the world becomes increasingly interconnected as a result of the Internet of Things (IoT) and smart gadgets, complex word identification may be essential in developing user interfaces and voice assistants that are intuitive to use and respond in a more human-like manner. The future of complex word identification has the potential to fundamentally alter how the researcher interact, learn, and communicate in a quickly changing the digital environment.

The future potential of Hindi lexical simplification holds enormous promise for boosting inclusion and accessibility in communication. There is a rising understanding of the significance of removing language barriers for a larger audience as technology and language processing continue to progress. Given its position as one of the most widely used languages in the world, Hindi stands to gain a lot from current initiatives towards vocabulary simplification. This area also has potential to be a part of the Make-in-India campaign, since the focus is on Hindi language.

Additionally, the fusion of artificial intelligence and natural language processing technology opens up new opportunities for cutting-edge Hindi lexical simplification applications. Advanced algorithms are able to examine the context, goal, and intended readership of a particular text, enabling more accurate and subtle simplification strategies. Furthermore, personalised and context-sensitive simplification models could be created to suit different reading abilities and preferences. The need for effective and efficient lexical simplification in Hindi will only increase as the world becomes more connected and information is shared across linguistic and cultural borders. This opens the door for collaborative initiatives between linguists, technologists, and educators to fully utilise Hindi's lexical simplification, creating a more welcoming and open digital environment for everyone.

To conclude, the research not only pushes the boundaries of linguistics and language technology, but also provides a practical means to increase communication accessibility and inclusion. The complex interaction between language nuances and technical innovation that underlies this research has the potential to alter how Hindi content is comprehended, transmitted, and valued. The impact of this work extends far beyond the boundaries of academic study, touching people with different backgrounds, cultural backgrounds, and linguistic proficiency as the digital age propels us into an era of global interconnection. Hindi complex word identification and lexical simplification are set for further development in the future, embracing simplicity as a beacon of efficient communication for future generations.

REFERENCES

1. A. Anula, “Tipos de textos, complejidad ling¨uıstica y facilicitaci´on lectora,” in *Actas del Sexto Congreso de Hispanistas de Asia, 2007*, pp. 45–61.
2. 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.
3. A. Sherman, Analytics of literature: A manual for the objective study of English prose and poetry. Ginn, 1893.
4. 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.
5. AbuRa'ed, A. G. T., & Saggion, H. (2018). LaSTUS/TALN at Complex Word Identification (CWI) 2018 Shared Task. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 159–65).
6. Achsan, H. T. Y., Suhartanto, H., Wibowo, W. C., Dewi, D. A., & Ismed, K. (2023). Automatic Extraction of Indonesian Stopwords. International Journal of Advanced Computer Science and Applications, 14(2).
7. Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez Agirre, A., Mihalcea, R., & Wiebe, J. (2016). Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *10th International Workshop on Semantic Evaluation, 2016* (pp. 497-511).
8. Agirre, E., De Lacalle, O. L., Fellbaum, C., Hsieh, S. K., Tesconi, M., Monachini, M., & Segers, R. (2010, July). Semeval-2010 task 17: All-words word sense disambiguation on a specific domain. In *Proceedings of the 5th international workshop on semantic evaluation* (pp. 75-80).
9. Alajmi, A., Saad, E. M., & Darwish, R. R. (2012). Toward an ARABIC stop-words list generation. International Journal of Computer Applications, 46(8), 8-13.
10. Alarcon, R., Moreno, L., Martínez, P., & Macías, J. A. (2024). EASIER system. Evaluating a Spanish lexical simplification proposal with people with cognitive impairments. International Journal of Human–Computer Interaction, 40(5), 1195-1209.
11. Alarcon, R., Moreno, L., & Martínez, P. (2023). EASIER corpus: A lexical simplification resource for people with cognitive impairments. Plos one, 18(4), e0283622.
12. Alu´ısio, S. and Gasperin, C. (2010). Fostering digital inclusion and accessibility: the porsimples project for simplification of portuguese texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas* (pp. 46–53).
13. Aroyehun, S. T., Angel, J., Alvarez, D. A. P., & Gelbukh, A. (2018, June). Complex word identification: Convolutional neural network vs. feature engineering. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications* (pp. 322-327).
14. Athaiya, A., Modi, D., & Pareek, G. (2018, October). A genetic algorithm based approach for Hindi word sense disambiguation. In *2018 3rd international conference on communication and electronics systems (ICCES)* (pp. 11-14). IEEE.

15. Baker, L. D., & McCallum, A. K. (1998). Distributional clustering of words for text classification. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 96-103).
16. Belder, J. D., & Moens, M. F. (2012, March). A dataset for the evaluation of lexical simplification. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 426-437). Springer, Berlin, Heidelberg.
17. Bevilacqua, M., & Navigli, R. (2020, July). Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 2854-2864).
18. Bhagoliwal, "Readability formulae: Their reliability, validity and applicability in hindi," Journal of Education and Psychology, vol. 19, no. 1, 1961.
19. Bhat, K., Ghumare, V., Khadake, S., & Gadade, H. D. (2022, June). Web Extension for Lexical Simplification of Text. In *2022 2nd International Conference on Intelligent Technologies (CONIT)* (pp. 1-7). IEEE.
20. Bhattacharyya, P., Pande, P., & L. Hindi WordNet LDC2008L02. Web Download. Philadelphia: Linguistic Data Consortium, 2008.
21. Billami, M. B., Francois, T., and Gala, N. (2018). Resyf: a french lexicon with ranked synonyms. In *Proceedings of the 27th International Conference on Computational Linguistics*, (pp. 2570–2581).
22. Bingel, J., Schluter, N., & Alonso, H. M. (2016, June). CoastalCPH at SemEval-2016 Task 11: The importance of designing your Neural Networks right. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 1028-1033).
23. Biran, O., Brody, S., & Elhadad, N. (2011, June). Putting it simply: a context-aware approach to lexical simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 496-501).
24. Blevins, T., & Zettlemoyer, L. (2020, July). Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1006-1017).
25. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the association for computational linguistics, *5*, 135-146.
26. Borah, P. P., Talukdar, G., & Baruah, A. (2014). Approaches for word sense disambiguation–A survey. International Journal of Recent Technology and Engineering, 3(1), 35-38.
27. Brants, T. (2006). Web 1t 5-gram version 1. http://www. ldc. upenn. edu/Catalog/CatalogEntry. jsp? catalogId= LDC2006T13.
28. Brown, K. (2005). *Encyclopedia of language and linguistics* (Vol. 1). Elsevier.
29. Brown, Q., Kim, E., Crowell, J., & Tse, T. (2005, November). A text corpora-based estimation of the familiarity of health terminology. In *International Symposium on Biological and Medical Data Analysis* (pp. 184-192). Springer, Berlin, Heidelberg.
30. Caplan, D. (1992). Language: Structure, processing, and disorders. The MIT Press.
31. Carroll, J. A., Minnen, G., Pearce, D., Canning, Y., Devlin, S., and Tait, J. (1999). Simplifying text for language-impaired readers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics* (pp. 269–270).
32. Carroll, J., Minnen, G., Canning, Y., Devlin, S., and Tait, J. (1998). Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the*

AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology, (pp. 7– 10). Citeseer.
33. Coleman and T. L. Liau (1975). "A computer readability formula designed for machine scoring." Journal of Applied Psychology, vol. 60, no. 2, pages 283–284.
34. Coster, W. and Kauchak, D. (2011). Simple english wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 665–669).
35. Crossley, S., Heintz, A., Choi, J. S., Batchelor, J., Karimi, M., & Malatinszky, A. (2023). A large-scaled corpus for assessing text readability. Behavior Research Methods, 55(2), 491-507.
36. Dale (1931). "A comparison of two word lists," Educational Research Bulletin, 484–489.
37. Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. Educational research bulletin, 37-54.
38. De Belder, J., & Moens, M. F. (2010). Text simplification for children. *In Proceedings of the SIGIR workshop on accessible search systems* (pp. 19-26). ACM; New York.
39. De Hertog, D., & Tack, A. (2018). Deep learning architecture for complex word identification. In *Thirteenth Workshop of Innovative Use of NLP for Building Educational Applications* (pp. 328-334). Association for Computational Linguistics (ACL); New Orleans, Louisiana.
40. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL-HLT).
41. Devlin, S. (1998). The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic databases*.
42. Douma, W. H. (1960). Readability of Dutch farm papers: a discussion and application of readability-formulas. Readability of Dutch farm papers: a discussion and application of readability-formulas, (17).
43. Dransfield, J. E., & McCall, W. A. (1925). A technique for teaching silent reading. Teachers College Record, 26(9), 740-752.
44. E. L. Thorndike and I. Lorge, "The teacher's word book of 30,000 words." 1944.
45. E. L. Thorndike, "The teacher's word book," 1921.
46. 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.
47. Edmonds, P. and Cotton, S. (2001). Senseval-2: overview. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems* (pp. 1–5).
48. Elhadad, N. (2006). Comprehending technical texts: Predicting and defining unfamiliar terms. In *AMIA annual symposium proceedings* (Vol. 2006, pp. 239). American Medical Informatics Association.
49. Ferrés, D., Saggion, H., & Guinovart, X. G. (2017, September). An Adaptable Lexical Simplification Architecture for Major Ibero-Romance Languages. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems* (pp. 40-47).
50. Fix, E., & Hodges, J. L. (1989). Discriminatory analysis. Nonparametric discrimination: Consistency properties. International Statistical Review/Revue Internationale de Statistique, 57(3), 238-247.
51. Fry, "A readability formula that saves time," Journal of reading, vol. 11, no. 7, pages 513–578, 1968.

52. Gale, W. A., Church, K., & Yarowsky, D. (1992). One sense per discourse. In *Proceedings of Speech and Natural Language* (pp. 233-237).
53. Garcia, R. A. (2022). *Lexical simplification for the systematic support of cognitive accessibility guidelines* (Doctoral dissertation, Universidad Carlos III de Madrid).
54. Gautam, C. B. S., & Sharma, D. K. (2016, August). Hindi word sense disambiguation using Lesk approach on bigram and trigram words. In *Proceedings of the International Conference on Advances in Information Communication Technology & Computing* (pp. 1-5).
55. Glavaš, G., & Štajner, S. (2021). A multi-task learning approach to lexical simplification. Information Processing & Management, 58(1), 102316.
56. Gonzalez-Agirre, A., Castillo, M., & Rigau, G. (2012, May). A proposal for improving WordNet Domains. In *Proceedings of the Language Resource and Evaluation Conference* (Vol. 3457).
57. Gooding, S. (2022, May). On the Ethical Considerations of Text Simplification. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)* (pp. 50-57).
58. Gooding, S., & Tragut, M. (2022). One Size Does Not Fit All: The Case for Personalised Word Complexity Models. *arXiv preprint arXiv:2205.02564*.
59. Gooding, S., Kochmar, E., Yimam, S. M., & Biemann, C. (2021, June). Word complexity is in the eye of the beholder. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4439-4449).
60. Gunning, R. (1952). The technique of clear writing.
61. H. Mc Laughlin, Smog grading-a new readability formula, Journal of reading, vol. 12, no. 8, pp. 639–646, 1969.
62. Hadiwinoto, C., Ng, H. T., & Gan, W. C. (2019, November). Improved Word Sense Disambiguation Using Pre-Trained Contextualized Word Representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 5297-5306).
63. Hao, L., & Hao, L. (2008, December). Automatic identification of stop words in chinese text classification. In *2008 International conference on computer science and software engineering* (Vol. 1, pp. 718-722). IEEE.
64. Hartmann, N. S., Paetzold, G. H., & Aluísio, S. M. (2018, September). SIMPLEX-PB: A lexical simplification database and benchmark for Portuguese. In *International Conference on Computational Processing of the Portuguese Language* (pp. 272-283). Springer, Cham.
65. Hauser, R., Vamvas, J., Ebling, S., & Volk, M. (2022, June). A multilingual simplified language news corpus. In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with Reading DIfficulties (READI) within the 13th Language Resources and Evaluation Conference* (pp. 25-30).
66. Hazawawi, M. Zakaria, and S. Hisham, Formulating an algorithm to detect readability level of malay texts. In *Proceedings of Mechanical Engineering Research Day 2017,* vol. 2017, pp. 77–78.
67. Hmida, F., Billami, M., Franc¸ois, T., and Gala, N. (2018). Assisted lexical simplification for french native children with reading difficulties. In *The Workshop of Automatic Text Adaptation, 11th International Conference on Natural Language Generation*.

68. Horn, C., Manduca, C., & Kauchak, D. (2014, June). Learning a lexical simplifier using wikipedia. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 458-463).
69. House, J. (2014). English as a global lingua franca: A threat to multilingual communication and translation? Language Teaching, 47(3):363–376.
70. Hsueh, S., Saggion, H., & Mille, S. (2012, May). Text simplification tools for spanish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 1665-1671).
71. Huang, E. H., Socher, R., Manning, C. D., & Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th annual meeting of the association for computational linguistics (Volume 1: Long papers)* (pp. 873-882).
72. Huang, L., Sun, C., Qiu, X., & Huang, X. J. (2019, November). GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3509-3514).
73. Iacobacci, I., Pilehvar, M. T., & Navigli, R. (2016, August). Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 897-907).
74. Ide, Y., Mita, M., Nohejl, A., Ouchi, H., & Watanabe, T. (2023, July). Japanese Lexical Complexity for Non-Native Readers: A New Dataset. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 477-487).
75. Jha, P., Agarwal, S., Abbas, A., & Siddiqui, T. J. (2023). A novel unsupervısed graph-based algorıthm for Hindi word sense disambiguation. SN Computer Science, 4(5), 675.
76. Jha, V., Manjunath, N., Shenoy, P. D., & Venugopal, K. R. (2016, January). Hsra: Hindi stopword removal algorithm. In *2016 International Conference on Microelectronics, Computing and Communications (MicroCom)* (pp. 1-5). IEEE.
77. Joshi, A., Patel, Y., Bhattacharya, P., & Carman, M. (2020). IndicBERT: A Pretrained Language Model for Indian Languages.
78. Joshi, H., Pareek, J., Patel, R., & Chauhan, K. (2012, December). To stop or not to stop—Experiments on stopword elimination for information retrieval of Gujarati text documents. In *2012 Nirma University International Conference on Engineering (NuiCONE)* (pp. 1- 4). IEEE.
79. Kaddoura, S., & Nassar, R. (2024). EnhancedBERT: A feature-rich ensemble model for Arabic word sense disambiguation with statistical analysis and optimized data collection. Journal of King Saud University-Computer and Information Sciences, 36(1), 101911.
80. Kajiwara, T., & Yamamoto, K. (2015, July). Evaluation dataset and system for Japanese lexical simplification. In *Proceedings of the ACL-IJCNLP 2015 Student Research Workshop* (pp. 35-40).
81. Karuppaiah, D., & Vincent, P. D. R. (2021). Word sense disambiguation in Tamil using Indo-WordNet and cross-language semantic similarity. International Journal of Intelligent Enterprise, 8(1), 62-73.
82. Kauchak, D. (2016). Pomona at semeval-2016 task 11: Predicting word complexity based on corpus frequency. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1047–1051.
83. Kaur, J., & Saini, J. R. (2016). POS Word Class Based Categorization of Gurmukhi Language Stemmed Stop Words. In *Proceedings of First International Conference on*

Information and Communication Technology for Intelligent Systems: Volume 2 (pp. 3-10). Springer, Cham.
84. Kaur, J., & Saini, J. R. (2016, March). Punjabi Stop Words: A Gurmukhi, Shahmukhi and Roman Scripted Chronicle. In *Proceedings of the ACM Symposium on Women in Research 2016* (pp. 32-37). ACM.
85. Keller, F., & Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. Computational linguistics, 29(3), 459-484.
86. Kenton, J. D. M. W. C., & Toutanova, L. K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT* (pp. 4171-4186).
87. Kouzis-Loukas, D. (2016). Learning scrapy. Livery Place: Packt Publishing.
88. Krippendorff, K. (2007). Computing Krippendorff's Alpha-Reliability. http://www. asc. upenn. edu/Krippendorff/.
89. Kucukyilmaz, T., & Akin, T. (2023, September). A Feature-based Approach on Automatic Stopword Detection. In *Intelligent Systems Conference* (pp. 51-67). Cham: Springer Nature Switzerland.
90. Kulkarni, D. S., & Rodd, S. F. (2022). Word Sense Disambiguation for Lexicon-based Sentiment Analysis in Hindi. Webology, *19*(1).
91. Kunchukuttan, A., Kakwani, D., Golla, S., Bhattacharyya, A., Khapra, M. M., & Kumar, P. (2020). Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages.
92. Kunchukuttan, A., Kakwani, D., Golla, S., Gokul N, C., Bhattacharyya, A., Khapra, M. M., & Kumar, P. (2020). AI4Bharat-IndicNLP Corpus: Monolingual Corpora and Word Embeddings for Indic Languages.
93. Lahmar, O., & Piras, L. (2023). Making sense and transparency in finance literature: Evidence from trends in readability. Research in International Business and Finance, 64, 101900.
94. Larkey, L. S., Connell, M. E., & Abduljaleel, N. (2003). Hindi CLIR in thirty days. ACM Transactions on Asian Language Information Processing (TALIP), 2(2), 130-142.
95. Lee, J. S. and Yeung, C. Y. (2018). Personalizing lexical simplification. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 224–232).
96. Lesk, M. (1986, June). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. *In Proceedings of the 5th annual international conference on Systems documentation* (pp. 24-26).
97. Likert, R. (1932). A technique for the measurement of attitudes. Archives of psychology.
98. Lo, R. T. W., He, B., & Ounis, I. (2005, January). Automatically building a stopword list for an information retrieval system. In Journal on Digital Information Management: Special Issue on the 5th DutchBelgian Information Retrieval Workshop (DIR) (Vol. 5, pp. 17-24).
99. Locke, J. (2003). The plain language movement. Journal-American Medical Writers Association, 18(1), 5-8.
100. Loureiro, D., & Jorge, A. (2019, July). Language Modelling Makes Sense: Propagating Representations through WordNet for Full-Coverage Word Sense Disambiguation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 5682-5691).
101. Maddela, M., & Xu, W. (2018, January). A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

102. Maddela, M., & Xu, W. (2018, January). A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.
103. Makrehchi, M., & Kamel, M. S. (2017). Extracting domain-specific stopwords for text classifiers. Intelligent Data Analysis, 21(1), 39-62.
104. Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. The annals of mathematical statistics, pages 50–60.
105. Mc Laughlin, G. H. (1969). SMOG grading-a new readability formula. Journal of reading, 12(8), 639-646.
106. McCall, W. A., & Crabbs, L. M. (1925). Standard Test Lessons in Reading. Teachers College Record, 27(3), 183-191.
107. McCarthy, D., & Navigli, R. (2009). The English lexical substitution task. Language resources and evaluation, 43(2), 139-159.
108. Meli, M., Tanti, M., & Porter, C. (2024, March). Towards Content Accessibility Through Lexical Simplification for Maltese as a Low-Resource Language. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion* (pp. 41-51).
109. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
110. Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. Language and cognitive processes, 6(1), 1-28.
111. Mishra, B. K., & Jain, S. (2023). An Innovative Method for Hindi Word Sense Disambiguation. SN Computer Science, 4(6), 704.
112. Mohammed, A., & Kora, R. (2022). An effective ensemble deep learning framework for text classification. Journal of King Saud University-Computer and Information Sciences, 34(10), 8825-8837.
113. Moreno, L., Alarcon, R., Segura-Bedmar, I., & Martínez, P. (2019, June). Lexical simplification approach to support the accessibility guidelines. In *Proceedings of the XX International Conference on Human Computer Interaction* (pp. 1-4).
114. Murphy, C. (2012). Effective Listings of Function Stop words for Twitter. International Journal of Advanced Computer Science and Applications, 3(6).
115. Murphy, K. P. (2006). Naive bayes classifiers. *University of British Columbia*, *18*(60), 1-8.
116. 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.
117. Na, D., & Xu, C. (2015). Automatically generation and evaluation of stop words list for Chinese patents. Telkomnika, 13(4), 1414.
118. Narayan, S., & Gardent, C. (2014, June). Hybrid simplification using deep semantics and machine translation. In *The 52nd annual meeting of the association for computational linguistics* (pp. 435-445).
119. Navigli, R. (2009). Word sense disambiguation: A survey. ACM computing surveys (CSUR), 41(2), 1-69.
120. Navigli, R., & Ponzetto, S. P. (2010, July). BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 216-225).
121. Nishihara, D., & Kajiwara, T. (2020, May). Word Complexity Estimation for Japanese Lexical Simplification. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 3114-3120).

122. North, K., Dmonte, A., Ranasinghe, T., Shardlow, M., & Zampieri, M. (2023, July). ALEXSIS+: Improving Substitute Generation and Selection for Lexical Simplification with Information Retrieval. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 404-413).
123. North, K., & Zampieri, M. (2023). Features of lexical complexity: insights from L1 and L2 speakers. Frontiers in Artificial Intelligence, 6.
124. North, K., Zampieri, M., & Ranasinghe, T. (2022, October). ALEXSIS-PT: A New Resource for Portuguese Lexical Simplification. In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 6057-6062).
125. Ogden, C. K. (1930). Basic english: A general introduction with rules and grammar.
126. Ortiz-Zambrano, J. A., Espin-Riofrio, C., & Montejo-Raéz, A. (2023). LegalEc: A New Corpus for Complex Word Identification Research in Law Studies in Ecuatorian Spanish. Procesamiento del Lenguaje Natural, 71, 247-259.
127. 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.
128. Paetzold, G. and Specia, L. (2016b). SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 560–569). Association for Computational Linguistics.
129. Paetzold, G. H., & Specia, L. (2016a). Unsupervised lexical simplification for non-native speakers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 1715-1725).
130. Paetzold, G., & Specia, L. (2016c, May). Benchmarking lexical simplification systems. *In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 3074-3080).
131. Pal, A. R., Saha, D., Dash, N. S., Naskar, S. K., & Pal, A. (2019). A novel approach to word sense disambiguation in Bengali language using supervised methodology. *Sādhanā*, *44*(8), 1-12.
132. Panjwani, R., Kanojia, D., & Bhattacharyya, P. (2018, January). pyiwn: a Python based API to access Indian language WordNets. In *Proceedings of the 9th Global Wordnet Conference* (pp. 378-383).
133. Pasini, T. (2021, January). The knowledge acquisition bottleneck problem in multilingual word sense disambiguation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence* (pp. 4936-4942).
134. Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).
135. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 2227-2237).
136. Qasmi, N. H., Zia, H. B., Athar, A., & Raza, A. A. (2020, May). SimplifyUR: unsupervised lexical text simplification for Urdu. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 3484-3489).

137. Qiang, J., Li, Y., Zhu, Y., Yuan, Y., & Wu, X. (2020). Lsbert: A simple framework for lexical simplification. *arXiv preprint arXiv:2006.14939*.
138. Quijada, M. and Medero, J. (2016). Hmc at semeval2016 task 11: Identifying complex words using depth-limited decision trees. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 1034–1037).
139. R. Brouwer, "Onderzoek naar de leesmoeilijkheden van nederlands proza," Pedagogische studi¨en, vol. 40, pp. 454–464, 1963.
140. 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.
141. 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.
142. R. McCallum and J. L. Peterson. (1982). Computer-based readability indexes, in *Proceedings of the ACM'82 Conference* (pp. 44–48).
143. R. Senter and E. A. Smith, "Automated readability index," CINCINNATI UNIV OH, Tech. Rep., 1967.
144. R. Weir and C. Ritchie. (2006). "Estimating readability with the strathclyde readability measure," in *ICT in the Analysis, Teaching and Learning of Languages, Preprints of the ICTATLL Workshop 2006* (pp. 25–32).
145. Raganato, A., Camacho-Collados, J., & Navigli, R. (2017, April). Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (pp. 99-110).
146. Rakholia, R. M., & Saini, J. R. (2016). Lexical classes based stop words categorization for Gujarati language. In *2016 2nd International Conference on Advances in Computing, Communication, & Automation (ICACCA)(Fall)* (pp. 1-5). IEEE.
147. Rani, R., & Lobiyal, D. K. (2018). Automatic Construction of Generic Stop Words List for Hindi Text. Procedia computer science, 132, 362- 370.
148. Raulji, J. K., & Saini, J. R. (2017, January). Generating Stopword List for Sanskrit Language. In *2017 IEEE 7th International Advance Computing Conference (IACC)* (pp. 799-802). IEEE.
149. Rello, L., Baeza-Yates, R., & Saggion, H. (2013). The impact of lexical simplification by verbal paraphrases for people with and without dyslexia. In *Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part II 14* (pp. 501-512). Springer Berlin Heidelberg.
150. Rouhizadeh, H., Shamsfard, M., Tajalli, V., & Rouhziadeh, M. (2021). Persian-WSD-Corpus: A Sense Annotated Corpus for Persian All-words Word Sense Disambiguation. *arXiv preprint arXiv:2107.01540*.
151. S. Stajner and H. Saggion. (2013) "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* (pp. 374–382).
152. Saeed, A., Nawab, R. M. A., Stevenson, M., & Rayson, P. (2019). A sense annotated corpus for all-words Urdu word sense disambiguation. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 18(4), 1-14.
153. Saeed, A., Nawab, R. M. A., Stevenson, M., & Rayson, P. (2019b). A word sense disambiguation corpus for Urdu. Language Resources and Evaluation, 53(3), 397-418.

154. Sagar, P., & Saini, J. R. (2023, December). A Three-Technique Pioneering Study for Unveiling Stopwords in Koshur Language. In *2023 11th International Conference on Intelligent Systems and Embedded Design (ISED)* (pp. 1-8). IEEE.
155. Saggion, H., Bott, S., & Rello, L. (2013). Comparing resources for Spanish lexical simplification. In *Statistical Language and Speech Processing: First International Conference, SLSP 2013, Tarragona, Spain*, July 29-31, 2013. Proceedings 1 (pp. 236-247). Springer Berlin Heidelberg.
156. Saini, J. R., & Rakholia, R. M. (2016). On continent and script-wise divisions-based statistical measures for stop-words lists of international languages. Procedia Computer Science, 89, 313-319.
157. Salton, G. & McGill, M. J. Introduction to Modern Information Retrieval. McGraw Hill Book Co., New York, 1983.
158. Sanjay, S. P., & Soman, K. P. (2016, June). Amritacen at semeval-2016 task 11: Complex word identification using word embedding. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 1022-1027).
159. Scarlini, B., Pasini, T., & Navigli, R. (2020a, April). Sensembert: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 05, pp. 8758-8765).
160. Scarlini, B., Pasini, T., & Navigli, R. (2020b, November). With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 3528-3539).
161. Shardlow, M. (2013, August). The cw corpus: A new resource for evaluating the identification of complex words. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations* (pp. 69-77).
162. Shardlow, M. (2014). A survey of automated text simplification. International Journal of Advanced Computer Science and Applications, 4(1):58–70.
163. Shardlow, M. (2015). A survey of automated lexical simplification. Journal of Artificial Intelligence Research, 53, 1-40.
164. Shardlow, M., Evans, R., & Zampieri, M. (2022). Predicting lexical complexity in English texts: the Complex 2.0 dataset. Language Resources and Evaluation, 1-42.
165. Shardlow, M., & Przybyła, P. (2023, September). Simplification by Lexical Deletion. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability* (pp. 44-50).
166. Sheang, K. C. (2019, September). Multilingual complex word identification: Convolutional neural networks with morphological and linguistic features. In *Proceedings of the Student Research Workshop Associated with RANLP 2019* (pp. 83-89).
167. Sheang, O. (2016, June). Ai-ku at semeval-2016 task 11: Word embeddings and substring features for complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 1042-1046).
168. Siddharthan, A. (2014). A survey of research on text simplification. ITL-International Journal of Applied Linguistics, 165(2):259–298.
169. Siddharthan, A., Nenkova, A., & McKeown, K. (2015). Syntactic simplification and text cohesion: Results of a large-scale study. Transactions of the Association for Computational Linguistics, 3, 193-206.
170. Silva, C., & Ribeiro, B. (2003). The importance of stop word removal on recall values in text categorization. In *Proceedings of the International Joint Conference on Neural Networks*, 2003. (Vol. 3, pp. 1661-1666). IEEE.

171. 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.
172. 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*.
173. Smokotin, V. M., Alekseyenko, A. S., and Petrova, G. I. (2014). The phenomenon of linguistic globalization: English as the global lingua franca (eglf). ProcediaSocial and Behavioral Sciences, 154:509–513
174. Soler, A. G., Apidianaki, M., & Allauzen, A. (2018). A comparative study of word embeddings and other features for lexical complexity detection in French. In *Actes de la Conférence TALN. Volume 1-Articles longs, articles courts de TALN* (pp. 499-508).
175. Song, J., Shen, Y., Lee, J., & Hao, T. (2020). A Hybrid Model for Community-Oriented Lexical Simplification. In *Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part I 9* (pp. 132-144). Springer International Publishing.
176. Specia, L., Jauhar, S. K., & Mihalcea, R. (2012). Semeval-2012 task 1: English lexical simplification. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)* (pp. 347-355).
177. Štajner, S. (2021). Automatic text simplification for social good: progress and challenges. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2637-2652.
178. Stajner, S., Saggion, H., and Ponzetto, S. P. (2019). Improving lexical coverage of text simplification systems for Spanish. Expert Systems with Applications, 118:80–91.
179. Tripathi, P., Mukherjee, P., Hendre, M., Godse, M., & Chakraborty, B. (2020). Word sense disambiguation in Hindi language using score based modified lesk algorithm. International Journal of Computing and Digital Systems, 10, 2-20.
180. Vajjala, S., Gupta, S., & Rama, T. (2018). An evaluation dataset for sentence simplification. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 2224-2234).
181. Valentini, M., Weber, J., Salcido, J., Wright, T., Colunga, E., & von der Wense, K. (2023, December). On the Automatic Generation and Simplification of Children’s Stories. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 3588-3598).
182. Venugopal-Wairagade, G., Saini, J. R., & Pramod, D. (2020). Novel Language Resources for Hindi: An Aesthetics Text Corpus and a Comprehensive Stop Lemma List. International Journal of Advanced Computer Science and Applications, 11(1).LR-2: The Open Parallel Corpus from http://opus.nlpl.eu/.
183. Vial, L., Lecouteux, B., & Schwab, D. (2019, July). Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. In *Proceedings of the 10th Global Wordnet Conference* (pp. 108-117).
184. W. S. Gray and B. E. Leary, “What makes a book readable.” 1935.
185. Walia, H., Rana, A., & Kansal, V. (2018a, August). Word sense disambiguation: Supervised program interpretation methodology for Punjabi language. In *2018 7th*

international conference on reliability, infocom technologies and optimization (Trends and future directions)(ICRITO) (pp. 762-767). IEEE.

186. Walia, H., Rana, A., & Kansal, V. (2018b, January). A supervised approach on Gurmukhi word sense disambiguation using K-NN method. In *2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 743-746). IEEE.
187. Wang, T., Chen, P., Rochford, J., & Qiang, J. (2016, March). Text simplification using neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 30, No. 1).
188. Wiedemann, G., Remus, S., Chawla, A., & Biemann, C. (2019). Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. *arXiv preprint arXiv:1909.10430*.
189. Wilson, M. (1988). MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior research methods, instruments, & computers*, *20*(1), 6-10.
190. Xu, Y., Wang, L., Wan, X., & Xiao, J. (2021). A Neural Machine Translation Based Framework for Syntactic Lexical Simplification. IEEE Access, 9, 46573-46583.
191. Yang, C. Z., Li, J. J., & Lin, S. C. (2023, October). Lexical Complexity Prediction using Word Embeddings. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)* (pp. 279-287).
192. Yang, Y., & Pedersen, J. O. (1997, July). A comparative study on feature selection in text categorization. In *Proceedings of the International Conference on Machine Learning* (Vol. 97, No. 412-420, pp. 35).
193. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*.
194. Yao, Z., & Ze-wen, C. (2011, March). Research on the construction and filter method of stop-word list in text preprocessing. In *2011 Fourth International Conference on Intelligent Computation Technology and Automation* (Vol. 1, pp. 217-221). IEEE.
195. Yarowsky, D. (1995, June). Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics* (pp. 189-196).
196. Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Štajner, S., . & Zampieri, M. (2018, June). A Report on the Complex Word Identification Shared Task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 66-78).
197. Yimam, S. M., Štajner, S., Riedl, M., & Biemann, C. (2017a, September). Multilingual and cross-lingual complex word identification. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP* (pp. 813-822).
198. Yimam, S. M., Štajner, S., Riedl, M., & Biemann, C. (2017b, November). CWIG3G2-complex word identification task across three text genres and two user groups. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 401-407).
199. Zampieri, M., Malmasi, S., Paetzold, G., & Specia, L. (2017). Complex word identification: Challenges in data annotation and system performance. *arXiv preprint arXiv:1710.04989*.
200. Zeman, D., Hajic, J., Popel, M., Potthast, M., Straka, M., Ginter, F., & Petrov, S. (2018, October). CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies* (pp. 1-21).

201. Zou, F., Wang, F. L., Deng, X., Han, S., & Wang, L. S. (2006, April). Automatic construction of Chinese stop word list. In *Proceedings of the 5th WSEAS international conference on Applied computer science* (pp. 1010-1015).

Appendices

Appendix A: List of Publications

Contributions in Journals

1. Venugopal-Wairagade, Gayatri, Saini, J. R., & Pramod, Dhanya (2020). Novel Language Resources for Hindi: An Aesthetics Text Corpus and a Comprehensive Stop Lemma List. *International Journal of Advanced Computer Science and Applications*, *11*(1), 233-239.
2. Venugopal, Gayatri, Pramod, Dhanya, & Saini, J. R. (2021). Analyzing Complex Words in Hindi using Parameters of Classical Readability Formulae (Part 1). *Computer Science Journal of Moldova*, *87*(3), 366-387.
3. Venugopal, Gayatri, Pramod, Dhanya, & Saini, J. R. (2022). Revisiting the role of classical readability formulae parameters in complex word identification (Part 2). *Computer Science Journal of Moldova*, *88*(1), 49-63.

Contributions in Conference Proceedings

1. Venugopal, Gayatri, Pramod, Dhanya, & Shekhar, R. (2022, June). CWID-hi: A dataset for complex word identification in Hindi text. In *Proceedings of the Thirteenth Language resources and evaluation Conference* (pp. 5627-5636).
2. Venugopal, Gayatri, & Pramod, Dhanya (2022, December). Bibliometric Analysis of Studies on Lexical Simplification. In *Proceedings of the International Conference on Hybrid Intelligent Systems* (pp. 3-12).

Popular Talks

Title of the talk: Lexical Simplification of Hindi Text

Conference: Women Who Code: CONNECT Forward Virtual Conference 2020

Link: https://www.youtube.com/watch?v=ly4VlKwPGLs

1. Venugopal-Wairagade, Gayatri, Saini, J. R., & Pramod, Dhanya (2020). Novel Language Resources for Hindi: An Aesthetics Text Corpus and a Comprehensive Stop Lemma List. *International Journal of Advanced Computer Science and Applications*, *11*(1), 233-239.

(IJACSA) International Journal of Advanced Computer Science and Applications,
Vol. 11, No. 1, 2020

Novel Language Resources for Hindi: An Aesthetics Text Corpus and a Comprehensive Stop Lemma List

Gayatri Venugopal-Wairagade[1], Jatinderkumar R. Saini[2]*, Dhanya Pramod[3]
Symbiosis Institute of Computer Studies and Research, Symbiosis International (Deemed University), Pune, India[1, 2]
Symbiosis Centre for Information Technology, Symbiosis International (Deemed University), Pune, India[3]

***Abstract*—This paper is an effort to complement the contributions made by researchers working toward the inclusion of non-English languages in natural language processing studies. Two novel Hindi language resources have been created and released for public consumption. The first resource is a corpus consisting of nearly thousand pre-processed fictional and non-fictional texts spanning over hundred years. The second resource is an exhaustive list of stop lemmas created from 12 corpora across multiple domains, consisting of over 13 million words, from which more than 200,000 lemmas were generated, and 11 publicly available stop word lists comprising over 1000 words, from which nearly 400 unique lemmas were generated. This research lays emphasis on the use of stop lemmas instead of stop words owing to the presence of various, but not all morphological forms of a word in stop word lists, as opposed to the presence of only the root form of the word, from which variations could be derived if required. It was also observed that stop lemmas were more consistent across multiple sources as compared to stop words. In order to generate a stop lemma list, the parts of speech of the lemmas were investigated but rejected as it was found that there was no significant correlation between the rank of a word in the frequency list and its part of speech. The stop lemma list was assessed using a comparative method. A formal evaluation method is suggested as future work arising from this study.**

***Keywords*—Hindi; corpus; aesthetics; stopwords; stoplemmas**

I. Introduction

One of the basic requirements to devise a tool to perform any task in Natural Language Processing (NLP) is a corpus that represents the target language or the target domain. Adhering to the 'Bender Rule', according to which researchers are required to name the language that was targeted by the study, we would like to inform the readers that the study focuses on Hindi, the official language of India. The language ranks third on the list of the languages with the largest number of first language speakers in the world [1]. The study aims at building a corpus in the aesthetics domain and utilizing the corpus, along with other corpora to publish an exhaustive list of stop lemmas based on their raw frequencies in the corpora. The corpus has been released under the GNU General Public License for public use, in order to ease the process of corpus acquisition. It is also necessary to mention that the text and the corpora that were acquired from various sources have been used solely for academic and research purpose. The corpus was created because of the difficulty that we faced while searching for and acquiring novels, stories and non-fictional content written by contemporary authors as well as content written by authors in India's pre-independence era, i.e., prior to 1947. The broad objective of the study is to utilize and release unbiased time-independent text in the form of a corpus. Since the study emphasizes the lexical aspect of text, features related to context and discourse have not been discussed in this paper. The stop lemma list created as part of this work has been built using text from multiple sources, thus making it suitable for generic consumption. The reported outcome is best as on date subject to the data used for this research. We believe that this corpus and the list of stop lemmas would be a useful resource for NLP tasks such as creating language models, text classification and information retrieval in Hindi.

The remaining paper is organized into three sections. Section II of the paper consists of the existing work in this area and the research questions. Section III contains the description of the corpus and the methodology along with the final list of stop lemmas. Section IV consists of the concluding remarks.

II. Existing Work and Objectives

Stop words are words that are present in a sentence solely for grammatical reasons and do not contribute to the information obtained from the text [2]. Hence if these stop words are identified and removed before using the text for a task, the performance of the task could improve. Many studies have focused on the importance of removing stop words as a pre-processing step for text processing tasks [3], [4] and [5]. Author in [6] manually extracted stop words based on parts of speech such as pronouns, prepositions, conjunctions etc. from two news-based corpora to create a stop word list in Hindi consisting of 275 words. Author in [7] created a stop word list by converting words to lemmas in a corpus of news articles consisting of 441,153 words. Author in [8] proposed a method for automatic stop word generation that created stop word lists that matched the top twenty stop word lists from four publicly available lists. Their corpus was based on news articles. Although we found studies that focused on automatic stop word generation [9], [10], [11], [12], [13] and others that published the stop word lists for public use [14], we could not find an exhaustive publicly available list of stop words in Hindi based on multiple corpora. Another problem with multiple lists based on one or two corpora is the inconsistency of the words in the lists [15]. Author in [16] manually created a stop word list consisting of 256 words from Punjabi poetry and articles. This list was brought down to 184 unique words by lemmatizing the words. We would like to emphasize here that the researchers lemmatized the words after identifying a word as a stop word as opposed to lemmatizing all the words and then identifying the stop words. [17] collated a list of stop words in numerous languages spanning multiple countries.

*Corresponding Author.

2. Venugopal, Gayatri, Pramod, Dhanya, & Saini, J. R. (2021). Analyzing Complex Words in Hindi using Parameters of Classical Readability Formulae (Part 1). *Computer Science Journal of Moldova*, *87*(3), 366-387.

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

3. Venugopal, Gayatri, Pramod, Dhanya, & Saini, J. R. (2022). Revisiting the role of classical readability formulae parameters in complex word identification (Part 2). *Computer Science Journal of Moldova*, *88*(1), 49-63.

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.

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

† Equal contribution

‡ Equal contribution

4. Venugopal, Gayatri, Pramod, Dhanya, & Shekhar, R. (2022, June). CWID-hi: A dataset for complex word identification in Hindi text. In *Proceedings of the Thirteenth Language resources and evaluation Conference* (pp. 5627-5636).

Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5627–5636
Marseille, 20-25 June 2022

CWID-hi: A Dataset for Complex Word Identification in Hindi Text

Gayatri Venugopal[◇], **Dhanya Pramod**[*], **Ravi Shekhar**[†]
[◇]Symbiosis Institute of Computer Studies and Research, Symbiosis International (Deemed University), Pune, India
[*]Symbiosis Centre for Information Technology, Symbiosis International (Deemed University), Pune, India
[†]Cognitive Science Research Group, Queen Mary University of London, London, UK
gayatri.venugopal@sicsr.ac.in, dhanya@scit.edu, r.shekhar@qmul.ac.uk

Abstract

Text simplification is a method for improving the accessibility of text by converting complex sentences into simple sentences. Multiple studies have been done to create datasets for text simplification. However, most of these datasets focus on high-resource languages only. In this work, we proposed a complex word dataset for Hindi, a language largely ignored in text simplification literature. We used various Hindi knowledge annotators for annotation to capture the annotator's language knowledge. Our analysis shows a significant difference between native and non-native annotators' perception of word complexity. We also built an automatic complex word classifier using a soft voting approach based on the predictions from tree-based ensemble classifiers. These models behave differently for annotations made by different categories of users, such as native and non-native speakers. Our dataset and analysis will help simplify Hindi text depending on the user's language understanding. The dataset is available at https://zenodo.org/record/5229160.

Keywords: lexical simplification, complex word identification, dataset, classification, Hindi

1. Introduction

The significance of accessible technology has increased manifold today for creating an inclusive environment. Accessibility not only helps improve the quality of life of persons with impairments but also benefits individuals without impairments (Henry, 2006). We can associate accessibility with physical devices as well as with the content we consume in our day-to-day lives. We can achieve accessibility of textual content by simplifying it. Text simplification refers to the process of modifying text in such a way that it becomes easier for the reader to comprehend it. It is a sub-field of natural language processing, that has proven to be useful to children (De Belder and Moens, 2010), readers with language impairments (Caplan, 1992; Carroll et al., 1999; Carroll et al., 1998), readers with low literacy levels (Candido Jr et al., 2009), and non-native speakers (Paetzold and Specia, 2016c). Conceptual simplification, elaborative modification, syntactic simplification, and lexical simplification are different ways by which we can simplify text (Siddharthan, 2014).

In this work, we focus on lexical simplification of Hindi text, i.e., the process of identifying complex words in a given text and substituting them with their simpler synonyms based on the context of the target complex word (Paetzold and Specia, 2017). The term 'complex' in this context does not indicate syntactic or morphological complexity. A complex word is defined as a word that is difficult to understand by the reader (Finnimore et al., 2019; Yimam et al., 2017a). Shardlow (2014) has demonstrated the steps in a lexical simplification pipeline, which is constituted of complex word identification, substitution generation, word sense disambiguation, and synonym ranking. However, there is no work related to lexical simplification in Hindi, which, according to Ethnologue[1], has the third-largest number of speakers in the world, after English and Mandarin Chinese. Moreover, Hindi is the official language of the government of India[2], and 43.63% people in the country are Hindi speakers, according to the last census conducted in India[3]. Since people's vocabulary varies according to their familiarity with the language, it is essential to produce and distribute content that all can understand. Manual simplification of content could be time-consuming and laborious; therefore, we need to use automatic text simplification approaches.

Soni et al. (2013) performed sentence simplification on Hindi text. They performed simplification by splitting the sentence into multiple sentences, which involved modification of the grammatical structure of the sentence. Mishra et al. (2014) also experimented with sentence splitting in order to improve the quality of translation from Hindi to English. Both these studies are instances of syntactic simplification of Hindi text. The area of text simplification, specifically lexical simplification, is unexplored for Hindi. Lexical simplification studies have gained popularity in various other languages such as English, French, Brazilian Portuguese, Spanish, to name a few (Paetzold, 2016; Lee and Yeung, 2018; Hmida et al., 2018; Aluísio and Gasperin, 2010; Štajner et al., 2019; Bott et al., 2012). However, certain resources used for simplification, such as the Simple English Wikipedia (Coster and Kauchak,

[1]https://www.ethnologue.com/guides/ethnologue200
[2]https://rajbhasha.gov.in/en/official-language-resolution-1968
[3]https://censusindia.gov.in/2011census/Language-2011/Statement-4.pdf

5. Venugopal, Gayatri, & Pramod, Dhanya (2022, December). Bibliometric Analysis of Studies on Lexical Simplification. In *Proceedings of the International Conference on Hybrid Intelligent Systems* (pp. 3-12).

SPRINGER LINK

Log in

Find a journal Publish with us Search Cart

International Conference on Hybrid Intelligent Systems

↳ HIS 2022: **Hybrid Intelligent Systems** pp 3–12 | Cite as

Home > Hybrid Intelligent Systems > Conference paper

Bibliometric Analysis of Studies on Lexical Simplification

Gayatri Venugopal & Dhanya Pramod

Conference paper | First Online: 25 May 2023

Part of the Lecture Notes in Networks and Systems book series (LNNS,volume 647)

Abstract

Text simplification is the process of improving the accessibility of text by modifying the text in such a way that it becomes easy for the reader to understand, while at the same time retaining the meaning of the text. Lexical simplification is a subpart of text simplification wherein the words in the text are replaced with their simpler synonyms. Our study aimed to examine the work done in the area of lexical simplification in various languages around the world. We conducted this study to ascertain the progress of the field over the years. We included articles from journals indexed in Scopus, Web of Science and the Association for Computational Linguistics (ACL) anthology. We analysed various attributes of the articles and observed that journal publications received a significantly larger number of citations as compared to conference publications. The need for simplification studies in languages besides English was one of the other major findings. Although we saw an increase in collaboration among authors, there is a need for more collaboration among authors from different countries, which presents an opportunity for conducting cross-lingual studies in this area. The observations reported in this paper indicate the growth of this specialised area of natural language processing, and also direct researchers' attention to the fact that there is a wide scope for conducting more diverse research in this area. The data used for this study is available on https://github.com/gayatrivenugopal/bibliometric_lexical_simplification.

Keywords

bibliometric study | lexical simplification | natural language processing

References

1. Rello, L., Baeza-Yates, R., Bott, S., Saggion, H.: Simplify or help? Text simplification strategies for people with dyslexia. In: Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility, pp. 1–10 (May, 2013)

Google Scholar

Appendix B: Sanction Letter Received from Symbiosis International (Deemed University) for conducting a Minor Research Project

SYMBIOSIS INTERNATIONAL (DEEMED UNIVERSITY)

(Established under section 3 of the UGC Act, 1956)
Re-accredited by NAAC with 'A' grade (3.58/4) | Awarded Category – I by UGC
Founder: Prof. Dr. S. B. Mujumdar, M. Sc., Ph. D. (Awarded Padma Bhushan and Padma Shri by President of India)

SIU/SCRI/MRP Approval/2019/1591

March 29th, 2019

To,
Dr. Gayatri Venugopal,
Symbiosis Institute of Computer Studies and Research,
Pune

Sub: Financial Assistance to Dr. Gayatri Venugopal, Assistant Professor, Symbiosis Institute of Computer Studies and Research for undertaking Minor Research Project (2018-2019) in Computer Studies for the Project entitled "Complex Word Identification in Hindi Sentences"

Dear Sir/Madam,

This is in reference to the proposal submitted by Dr. Gayatri Venugopal, PI and Dr. Dhanya Pramod, Co-PI to Symbiosis Centre for Research and Innovation. The research proposals were invited for minor research projects with an aim to institutionalize research activities and promote excellence in research by providing financial support vide letter /email No. SIU/SCRI/Minor Research Project/2018-19 dated 1st Oct. 2018. The proposals were reviewed double blindly by external experts. Based on the recommendations and in accordance with the guidelines, the Symbiosis International (Deemed University) has approved an amount of Rs. 1,50,000/- as mentioned below:

Items	Sanctioned Amount
Recurring	
• Field work & travel	Rs. 25,000/-
• Manpower	Rs. 1,15,000/-
• Contingencies and Stationery	Rs. 10,000/
Total	Rs. 1,50,000/-

On receipt of the Approval letter, the Principal Investigator should inform the undersigned of his/her consent to implement the project and send the Acceptance Certificate attached herewith within a week, otherwise it will be presumed that the Principal Investigator (PI) is not willing to implement the project and the approval will be withdrawn.

The grant is subject to the terms and conditions as per the guidelines of Symbiosis International University and availability of funds.

Thanking you,

29/3/19

Dr. Dipak Tatpuje
Head - Research Projects,
Symbiosis Centre for Research and Innovation (SCRI), SIU

29/3/2019

Dr. Urvashi Rathod
Director – SCRI , SIU

Copy to: 1. Vice- Chancellor, SIU
2. Finance Officer, SIU
3. Dean, FoCS and Director, SICSR

Appendix C: Approval Received from the Independent Ethics Committee, Symbiosis International (Deemed University)

SYMBIOSIS INTERNATIONAL (DEEMED UNIVERSITY)

(Established under section 3 of the UGC Act. 1956)
Re-accredited by NAAC with 'A' grade

INDEPENDENT ETHICS COMMITTEE

(Registered with DCGI with registration number ECR/147/Indt/MH/2014/RR-17 dated 6th April 2018)

Dr. Raman Gangakhedkar
Chairman

Dr. Vasant Padbidri
Member

SIU/IEC/Admin/

Date: 9 July 2019

SYMBIOSIS INTERNATIONAL (DEEMED UNIVERSITY)

(Established under section 3 of the UGC Act. 1956)
Re-accredited by NAAC with 'A' grade

INDEPENDENT ETHICS COMMITTEE

(Registered with DCGI with registration number ECR/147/Indt/MH/2014/RR-17 dated 6th April 2018)

Dr. Raman Gangakhedkar
Chairman

SIU/IEC/Admin/

Dr. Vasant Padbidri
Member

Dr. Ravindra Ghooi
Member

Dr. Rajiv Yeravdekar
Member

Dr. Milind Telang
Member

Dr. Shailaja Kale
Member

Dr. Shashikala Gurpur
Member

Ms Ritu Chhabria
Member

Dr. Kuruvilla Pandikattu
Member

Mr. Umesh Isalkar
Member

Dr. Sonopant Joshi
Member Secretary

As the compliance is found satisfactory by the committee, your research proposal is herewith approved.

Prof. (Dr.) Sonopant Joshi
Member Secretary
Independent Ethics Committee- SIU

Dr. Joshi S. G.
Member Secretary
Independent Ethics Committee
SIU-Pune

Symbiosis International (Deemed University) IEC-SIU PUNE

Gram Lavale, Tal: Mulshi, Dist:Pune 412115Tel: +91-20-39116200/39116208/39116209 Fax No : +91-20-39116206

Appendix D: Minor Research Project Completion Certificate

SYMBIOSIS INTERNATIONAL (DEEMED UNIVERSITY)

(Established under section 3 of the UGC Act 1956)

Re-Accredited by NAAC with 'A' grade (3.58/4) I Awarded Category - I by UGC

Founder : Prof. Dr. S. B. Mujumdar M.Sc. Ph.D. (Awarded Padma Bhushan and Padma Shri by President of India)

SIU/SCRI/MRP/2021/2556 5, July 2021

Certificate for Completion of Minor Research Project

This is to certify that the Minor Research Project was awarded to **Gayatri Venugopal, Assistant Professor, Symbiosis Institute of Computer Studies and Research, Pune & Co-PI Dr. Dhanya Pramod, Professor and Director, Symbiosis Centre for Information Technology, Pune**

The project related information is as Follows:

Letter No. : - SIU/ SCRI/ Minor Research Project Approval/2019/1591 dated 29th March 2019.

Title of the Project : - Complex Word Identification in Hindi Sentences

Amount Sanctioned :- **Rs.1,50,000/-**

We have received the report on the work done in the project and its outcome. The research work has been assessed and found satisfactory. The project has been successfully completed. **The closure certificate will be issued after clearance from finance department of SIU**.

With Best Regards,

Dr. Dipak Tatpuje, Ph.D.
Head, Research Project, Symbiosis Centre for Research & Innovation (SCRI)
Symbiosis International (Deemed University)

Dr. Vinaykumar Rale, Ph. D.
Officiating Director, Symbiosis Centre for Research & Innovation (SCRI)
Symbiosis International (Deemed University

Lavale, Pune 412 115, Maharashtra, India. | Tel: +91-20-28116200/08/09 | Fax : +91-20-28116206
Email: info@siu.edu.in | Web: www.siu.edu.in

Appendix E: Subject Information Document

you. After you have completed the task of annotating 100 sentences, you would be presented with a set of words on the screen. You are required to rank the words according to their complexity, 1 being the least complex. Multiple such screens consisting of sets of words would be displayed.

The overall duration of the experiment would be approximately 90 minutes per participant.

Why have I been chosen?
You have been chosen because the study targets native as well as non-native Hindi speakers in the age group of 18-30 years.

Will I be paid for taking part?
Yes, you will be paid 1000 INR for taking part and completing all the tasks.

Do I have to take part?
Your participation is completely voluntary. You may choose not to take part at all.

Will my taking part in this trial be kept confidential?
Every effort will be made to keep the data confidential. The information provided by the participant during this research study will be kept confidential with the exception of potentially sharing the pictures, if any taken, and screen recordings for research purposes. Research records will be stored securely. Symbiosis International (Deemed University) employees may access or inspect records pertaining to this research as part of routine oversight or university business. The video recordings (if any) made in this study may also be used in presentations about this research. However, Symbiosis International (Deemed University) doctoral review committee/research advisory committee may inspect the participation records pertaining to this research. Some of these records could contain information that personally identifies the participant, i.e., the information that was entered in the sheet given to him/her.

Are there possible disadvantages and/or risks in taking part?
There are no disadvantages/risks of participating in this study.

What will happen to the results of the research project?
The data thus collected will be used to identify the features of a complex word. These features would be used to train a machine to automatically identify the complex words in a given Hindi sentence.

Ethical review of the study
The project has received ethical approval from the Independent Ethics Committee of Symbiosis International (Deemed University). (This line would be added only if the IEC of SIU approves the study).

Appendix F: Informed Consent Form

Study contacts: We would like you to ask us questions if there is anything about the study that you do not understand. You can call us at +91-9665856569 or email us at gayatri.venugopal@sicsr.ac.in or dhanya@scit.edu.

You can also contact the Independent Ethics Committee at the Symbiosis International (Deemed University) with any concerns that you have about your rights or welfare as a participant. This office can be reached at by email at iec@siu.edu.in.

Signatures: Your signature indicates that this study has been explained to you, that your questions have been answered, and that you agree to take part in this study.

- I confirm that I have read and understood the Subject Information Document and the Informed Consent Form.
- I have had the opportunity to ask questions and had them answered.
- I understand that all personal information will remain confidential and that all efforts will be made to ensure I cannot be identified.
- I agree that data gathered in this study may be stored anonymously and securely.
- I understand that my participation is voluntary.
- I have normal or corrected to normal vision.
- I do not have any reading disability to the best of my knowledge.
- I agree to take part in this study.

Name of Participant	Signature of Participant	Date
Name of Principal Investigator	Signature of Principal Investigator	Date

Appendix G: Data Collection Form

Data Collection Forms
I Participant Details Form

This form ascertains demographic data and your experience with Hindi.

1. Email ID: _____
2. Year of birth : _____
3. Gender: Male Female Others
4. Which state are you currently residing in? _____
5. What is your current occupation (if you are studying currently, please write 'student')

   _____
6. What is your mother tongue? _____
7. List 2-3 languages in which you are most comfortable to read, starting from your most comfortable language.

   _____
8. Which state of India have you lived in for the maximum number of years?

Also mention the number of years. _____
9. How often do you read content written in Hindi?
   a. I do not read content written in Hindi unless there is no alternative
   b. Rarely
   c. Once or twice a month
   d. Weekly
   e. Daily
10. If you read content written in Hindi, what kind of content do you read?
    a. Novels and short stories
    b. Non-fictional books
    c. Magazines
    d. News
    e. Poems
    f. Textbooks
    g. Official documents
    h. Others (please specify) _____
11. How many years did you study Hindi as part of the school curriculum? _____
12. If you studied Hindi as part of the school curriculum, which board/s was/were your school/s affiliated with? _____
13. If you studied Hindi as part of the school curriculum, mention the state/s in which you were residing at the time of studying, and also the number of years spent in the state/s.

_____

14. How many years did you study Hindi as part of the college curriculum? _____
15. Have you pursued/Are you pursuing higher education in Hindi?

    Yes No
16. If yes, please specify the degree/name of the programme. _____
17. Have you completed any certification course in Hindi? _____

    Yes No
18. If yes, please mention the name of the course and the duration.

    _____

I, _____ (participant name), hereby confirm that the details furnished above are true and correct to the best of my knowledge.

Participant ID: _____

Name of Participant: _____

Signature of Participant: _____

Appendix H: Language Resources used in the Study

LR-1: Wictionary Top 1900. (June, 2018) from https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/ Hindi_1900.

LR-2: 1000 Most Common Hindi Words from https://1000mostcommonwords.com/1000-most-common-hindi-words.

LR-3: First 100 High Frequency Words in Hindi. (May, 2012). In Hindi Language Blog from https://blogs.transparent.com/hindi/first-100-high-frequency-words-inhindi.

LR-4: Top 1000 Words from http://home.iitk.ac.in/~prasant/HindiCorpus/word.html.

LR-5: Piotr, Most Common Words by Language (2019) from https://github.com/oprogramador/mostcommon-words-by-language.

LR-6: Savand, A, Stop Words (2017) from https://github.com/Alir3z4/stop-words.

LR-7: Stopwords ISO, Stop Words-hi (2016) from https://github.com/stopwords-iso/stopwords-hi.

LR-8: Champion, Jason, Extra Stop Words (2016) from https://github.com/Xangis/extra-stopwords.

LR-9: Jha, Vandana, N, Manjunath, Shenoy, P Deepa, & K R, Venugopal (2018), "Hindi Language Stop Words List", Mendeley Data, v1 http://dx.doi.org/10.17632/bsr3frvvjc.

LR-10: Hindi Stopwords from https://www.ranks.nl/stopwords/hindi.

LR-11: Wikimedia Downloads from https://dumps.wikimedia.org/backup-index.html.

LR-12: Venugopal-Wairagade, G., Saini, J. R., & Pramod, D. (2020). Novel Language Resources for Hindi: An Aesthetics Text Corpus and a Comprehensive Stop Lemma List. International Journal of Advanced Computer Science and Applications, 11(1).

LR-13: The Open Parallel Corpus from http://opus.nlpl.eu/.

LR-14: CFILT Hindi Corpus from https://www.cfilt.iitb.ac.in/Downloads.html.

LR-15: Kunchukuttan, A., Mehta, P. & Bhattacharyya, P. (2018). The IIT Bombay English-Hindi Parallel Corpus. Language Resources and Evaluation Conference.

LR-16: English-Hindi Tourism Text Corpus – EILMT (October, 2016). EILMT Consortia, CDAC Pune from http://www.tdil-dc.in.

LR-17: Hindi-English Agriculture & Entertainment Text Corpus ILCI-II (May, 2017). ILCI Consortium, JNU from http://www.tdil-dc.in.

LR-18: Hindi Monolingual Text Corpus ILCI-II (June, 2017). ILCI-II, JNU from http://www.tdil-dc.in.

LR-19 Hindi-English Health Text Corpus-ILCI (April, 2012). ILCI Consortium, JNU. From http://www.tdil-dc.in.

Appendix I: Links to Public Repositories containing the Data and the Code used in the Study

The following resources have been released under an open source license each.

1. https://zenodo.org/record/5229160 - CWID-hi Dataset created as part of the Study
2. https://github.com/gayatrivenugopal/Hindi-Aesthetics-Corpus - Aesthetics Corpus created as part of the Study
3. https://github.com/gayatrivenugopal/hindi-corpus-stoplemmas - Exhaustive stop lemma list obtained from Hindi Text
4. https://github.com/gayatrivenugopal/ablation-study - Code to Conduct an Ablation Study for a Classification Task
5. https://github.com/gayatrivenugopal/Hindi-Lexical-Simplification - Data used in the Study, and the Code used for Pre-Processing the Data
6. https://github.com/gayatrivenugopal/complex-word-identification-hindi - Data and Code used for Complex Word Identification
7. https://github.com/gayatrivenugopal/bibliometric_lexical_simplification - Data used for Bibliometric Analysis
8. https://github.com/gayatrivenugopal/acl-anthology-browser - Code for extracting information using a keyword-based filter from the ACL Anthology dataset

Appendix J: Proforma 3

Proforma-3

Undertaking from the Ph.D. student while submitting his/her final Thesis to SCRI

Ref. No. _____

I, the undersigned, hereby declare and give an undertaking that the Ph.D. Thesis entitled Automated Lexical Simplification of Hindi Text under the Faculty of Computer Studies has been checked for its Similarity Index for Plagiarism through an authentic Turnitin software tool; and that the document has been prepared by me and is my original work and free of any plagiarism. It was found that:

1.	The Similarity Index (SI) was: *(Note: SI range: 0 to 10%; if SI is >10%, then the student cannot submit his/her thesis; attachment of SI report is mandatory)*	10%
2.	The ethical clearance for research work conducted obtained from: *(Note: Name the consent obtaining body; if 'not applicable' then write so)*	Independent Ethics Committee, SIU
3.	The material (adopted text, tables, figures, graphs, etc.) as has been obtained from other sources, has been duly acknowledged in the thesis: *(Note: Tick □ whichever is applicable)*	Yes / No

In case if any of the above-furnished information is found false at any point in time, then the University authorities can take action as deemed fit against me.

Date: _____

Place: _____

Signature, Full Name & PRN
of the Ph.D. student

Signature & Name of Supervisor

Appendix K: Similarity Report

001-Automated Lexical Simplification of Hindi Text

ORIGINALITY REPORT

10%	9%	5%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.lrec-conf.org Internet Source	2%
2	ibn.idsi.md Internet Source	2%
3	arxiv.org Internet Source	2%
4	www.math.md Internet Source	<1%
5	export.arxiv.org Internet Source	<1%
6	ebin.pub Internet Source	<1%
7	"Advances and Trends in Artificial Intelligence. Theory and Practices in Artificial Intelligence", Springer Science and Business Media LLC, 2022 Publication	<1%
8	aclanthology.org Internet Source	<1%

9	etheses.whiterose.ac.uk Internet Source	<1%
10	"Hybrid Intelligent Systems", Springer Science and Business Media LLC, 2023 Publication	<1%
11	preview.aclanthology.org Internet Source	<1%
12	"Word Sense Disambiguation", Springer Nature, 2007 Publication	<1%
13	"Computational Linguistics and Intelligent Text Processing", Springer Science and Business Media LLC, 2018 Publication	<1%
14	Prajna Jha, Shreya Agarwal, Ali Abbas, Tanveer J. Siddiqui. "A Novel Unsupervised Graph-Based Algorithm for Hindi Word Sense Disambiguation", SN Computer Science, 2023 Publication	<1%
15	www.ijimai.org Internet Source	<1%
16	Bong-Jun Yi, Do-Gil Lee, Hae-Chang Rim. "An All-Words Sense Tagging Method for Resource-Deficient Languages", Digital Scholarship in the Humanities, 2016 Publication	<1%

17	dc.etsu.edu Internet Source	<1%
18	fastercapital.com Internet Source	<1%
19	Andras Csomai. "Wikify!", Proceedings of the sixteenth ACM conference on Conference on information and knowledge management - CIKM 07 CIKM 07, 2007 Publication	<1%
20	kipdf.com Internet Source	<1%
21	www.frontiersin.org Internet Source	<1%
22	www.revistaocnos.com Internet Source	<1%
23	Ernest Kwame Ampomah, Zhiguang Qin, Gabriel Nyame. "Evaluation of Tree-Based Ensemble Machine Learning Models in Predicting Stock Price Direction of Movement", Information, 2020 Publication	<1%
24	Diego Kozlowski, Elisa Lannelongue, Frédéric Saudemont, Farah Benamara et al. "A three-level classification of French tweets in ecological crises", Information Processing & Management, 2020	<1%

	Publication	
25	Ton Duc Thang University Publication	<1 %
26	acikerisim.deu.edu.tr Internet Source	<1 %
27	diva-portal.org Internet Source	<1 %
28	www.i-asem.org Internet Source	<1 %
29	www.winlp.org Internet Source	<1 %
30	Horacio Saggion. "Automatic Text Simplification", Springer Science and Business Media LLC, 2017 Publication	<1 %
31	core.ac.uk Internet Source	<1 %
32	www.mdpi.com Internet Source	<1 %
33	"Trends and Perspectives in Linear Statistical Inference", Springer Science and Business Media LLC, 2018 Publication	<1 %
34	Kai North, Marcos Zampieri, Matthew Shardlow. "Lexical Complexity Prediction: An	<1 %

Overview", ACM Computing Surveys, 2023
Publication

35	www.ncbi.nlm.nih.gov Internet Source	<1%
36	www.tdx.cat Internet Source	<1%
37	Gayatri Venugopal-Wairagade, Jatinderkumar R., Dhanya Pramod. "Novel Language Resources for Hindi: An Aesthetics Text Corpus and a Comprehensive Stop Lemma List", International Journal of Advanced Computer Science and Applications, 2020 Publication	<1%
38	escholarship.org Internet Source	<1%
39	fdocuments.in Internet Source	<1%
40	Mora Pablo, Irasema. "The 'Native Speaker' Spin: The Construction of the English Teacher at a Language Department at a University in Central Mexico.", Canterbury Christ Church University (United Kingdom), 2020 Publication	<1%
41	Singh, Shekhar. "Facial Expression Recognition Using Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) for Data Augmentation and Image	<1%

	Generation", University of Nevada, Las Vegas, 2024 Publication	
42	journals.plos.org Internet Source	<1%
43	pure.tue.nl Internet Source	<1%
44	Aryan Nikul Patel, Ramalingam Murugan, Gautam Srivastava, Praveen Kumar Reddy Maddikunta et al. "An explainable transfer learning framework for multi-classification of lung diseases in chest X-rays", Alexandria Engineering Journal, 2024 Publication	<1%
45	Hisashi Kashima, Naoki Abe. "A Parameterized Probabilistic Model of Network Evolution for Supervised Link Prediction", Sixth International Conference on Data Mining (ICDM'06), 2006 Publication	<1%
46	Loureiro, Daniel. "Learning Word Sense Representations from Neural Language Models", Universidade do Porto (Portugal), 2024 Publication	<1%
47	clausiuspress.com Internet Source	<1%

48	ir.lib.uwo.ca Internet Source	<1%
49	Rafael Berlanga, Victoria Nebot, María Pérez. "Tailored semantic annotation for semantic search", Journal of Web Semantics, 2015 Publication	<1%
50	bmcbioinformatics.biomedcentral.com Internet Source	<1%
51	egusphere.copernicus.org Internet Source	<1%
52	livrepository.liverpool.ac.uk Internet Source	<1%
53	repository.library.carleton.ca Internet Source	<1%
54	www.irjet.net Internet Source	<1%
55	Hossein Rouhizadeh, Mehrnoush Shamsfard, Masoud Rouhizadeh. "Knowledge Based Word Sense Disambiguation with Distributional Semantic Expansion for the Persian Language", 2020 10th International Conference on Computer and Knowledge Engineering (ICCKE), 2020 Publication	<1%
56	Lecture Notes in Computer Science, 2015. Publication	<1%

57	Matthew Shardlow, Richard Evans, Marcos Zampieri. "Predicting lexical complexity in English texts: the Complex 2.0 dataset", Language Resources and Evaluation, 2022 Publication	<1%
58	Michael Stewart, Wei Liu. "E2EET: from pipeline to end-to-end entity typing via transformer-based embeddings", Knowledge and Information Systems, 2021 Publication	<1%
59	dokumen.pub Internet Source	<1%
60	researchonline.gcu.ac.uk Internet Source	<1%
61	"Computational Processing of the Portuguese Language", Springer Science and Business Media LLC, 2020 Publication	<1%
62	Haoran Wu, Fazhi He, Tongzhen Si, Yansong Duan, Xiaohu Yan. "HIGSA: Human image generation with self-attention", Advanced Engineering Informatics, 2023 Publication	<1%
63	Lindbloom, Cody James. "An Exploratory Analysis on the Lived Experiences of First-Year Students Participating in Living Learning	<1%

	Communities on a College Campus", University of Arkansas, 2023 Publication	
64	Michalis Papakostas, Kais Riani, Andrew Brian Gasiorowski, Yan Sun, Mohamed Abouelenien, Rada Mihalcea, Mihai Burzo. "Understanding Driving Distractions: A Multimodal Analysis on Distraction Characterization", 26th International Conference on Intelligent User Interfaces, 2021 Publication	<1 %
65	Suethanapornkul, Sakol. "Statistical Learning of Predictive Dependencies in the Tense-aspect System of a Miniature Language by English and Thai First Language Adults.", Georgetown University, 2020 Publication	<1 %
66	Williams, Anwen. "The Use of Herd Data to Teach Dairy Cattle Breeding in Further and Higher Education Colleges", Bangor University (United Kingdom), 2023 Publication	<1 %
67	kb.psu.ac.th Internet Source	<1 %
68	www.researchgate.net Internet Source	<1 %

69	"Brain Informatics", Springer Science and Business Media LLC, 2018 Publication	<1%
70	"Computational Processing of the Portuguese Language", Springer Science and Business Media LLC, 2018 Publication	<1%
71	"Testing the effect of synchronous speech tasks in the production of L2 speech rhythm in learners of Spanish as a second language.", Pontificia Universidad Catolica de Chile, 2020 Publication	<1%
72	A. Kralisch, B. Berendt. "Language-sensitive search behaviour and the role of domain knowledge", New Review of Hypermedia and Multimedia, 2005 Publication	<1%
73	Benzir Md Ahmed, Mohammed Eunus Ali, Mohammad Mehedy Masud, Mohammad Raihan Azad, Mahmuda Naznin. "After-meal blood glucose level prediction for type-2 diabetic patients", Heliyon, 2024 Publication	<1%
74	Casper Peeters, Koen Vijverberg, Marianne Pouwer, Bart Westerman, Maikel Boot, Suzan Verberne. "Evaluation of SURUS: a Named Entity Recognition System to Extract	<1%

Knowledge from Interventional Study Records", Cold Spring Harbor Laboratory, 2024
Publication

75	Joseph Paul Stemberger, Barbara Handford Bernhardt. "Handbook of Phonological Development", Brill, 1997 Publication	<1%
76	Kawintiranon, Kornraphop. "Detecting and Understanding of Information Pollution on Social Media", Georgetown University, 2023 Publication	<1%
77	Min Seok Lee, Seok Woo Yang, Hong Joo Lee. "Weight attention layer based document classification by incorporating information gain", Expert Systems, 2021 Publication	<1%
78	Viktar Atliha. "Improving image captioning methods using machine learning approaches", Vilnius Gediminas Technical University, 2023 Publication	<1%
79	Wutthipong Kongburan, Praisan Padungweang, Worarat Krathu, Jonathan H. Chan. "Enhancing metabolic event extraction performance with multitask learning concept", Journal of Biomedical Informatics, 2019	<1%

	Publication	
80	Ying Li. "Semantic Analysis for Topical Segmentation of Videos", c, 09/2007 Publication	<1 %
81	eprints.usm.my Internet Source	<1 %
82	koreascience.kr Internet Source	<1 %
83	medinform.jmir.org Internet Source	<1 %
84	mspace.lib.umanitoba.ca Internet Source	<1 %
85	pdffox.com Internet Source	<1 %
86	www.getit01.com Internet Source	<1 %
87	www.repository.cam.ac.uk Internet Source	<1 %
88	www.semanticscholar.org Internet Source	<1 %
89	Theory and Applications of Natural Language Processing, 2014. Publication	<1 %
90	Salma Y. Y. Hamad, Tao Ma, Constantinos Antoniou. "Analysis and Prediction of Bikesharing Traffic Flow – Citi Bike, New York", 2021 7th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), 2021 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		