



**CLASSIFICATION OF ENGLISH TEXTS '  
DIFFICULTY LEVELS THROUGH MACHINE  
LEARNING ALGORITHMS: A PROGRAM  
RECOMMENDATION**

**Somayyeh SHABESTANI**

**Master's Thesis**

**Department of Foreign Languages Teaching**

**Assoc. Prof. Dr. Merve GEÇİKLİ**

**2023**

**(All rights reserved.)**

T.C.  
ATATÜRK ÜNİVERSİTESİ  
EĞİTİM BİLİMLERİ ENSTİTÜSÜ  
YABANCI DİLLER EĞİTİMİ ANA BİLİM DALI  
İNGİLİZ DİLİ EĞİTİMİ BİLİM DALI

**MAKİNA ÖĞRENİM ALGORİTMALARI İLE İNGİLİZCE METİN ZORLUK  
SEVİYELERİNİN SINIFLANDIRILMASI: BİR YAZILIM ÖNERİSİ**

(Classification of English Texts' Difficulty Levels through Machine Learning Algorithms: A  
Program Recommendation)

YÜKSEK LİSANS TEZİ

Somayyeh SHABESTANI

Danışman: Doç. Dr. Merve GEÇİKLİ

Erzurum  
Haziran, 2023

## KABUL VE ONAY TUTANAĐI

Somayyeh SHABESTANI tarafından hazırlanan “Makina Öğrenim Algoritmaları İle İngilizce Metin Zorluk savunma sınavı sonucunda başarılı bulunarak jürimiz tarafından Yabancı Diller Eğitimi Ana Bilim Dalı, İngiliz Dili Eğitimi Bilim Dalında Yüksek Lisans tezi olarak kabul edilmiştir.

|               |                                                                        |                      |
|---------------|------------------------------------------------------------------------|----------------------|
| Jüri Başkanı: | Dr. Öğrt. Üyesi Zennure<br>ELGÜN GÜNDÜZ<br><i>Ardahan Üniversitesi</i> | Aslı Islak İmzalıdır |
| Danışman:     | Doç.Dr.Merve GEÇİKLİ<br><i>Atatürk Üniversitesi</i>                    | Aslı Islak İmzalıdır |
| Jüri Üyesi:   | Dr. Öğrt. Üyesi Veysel KARSLI<br><i>Atatürk Üniversitesi</i>           | Aslı Islak İmzalıdır |

Bu tezin Atatürk Üniversitesi Lisansüstü Eğitim ve Öğretim Yönetmeliđi'nin ilgili maddelerinde belirtilen şartları yerine getirdiđini onaylarım.

.... / ..... / 2023..

Prof. Dr. Adnan KÜÇÜKOĐLU

Enstitü Müdürü

Aslı Islak İmzalıdır

## ETİK VE BİLDİRİM SAYFASI

Yüksek Lisans Tezi olarak sunduğum “Makina Öğrenim Algoritmaları İle İngilizce Metin Zorluk Seviyelerinin Sınıflandırılması : Bir Yazılım Önerisi” başlıklı çalışmanın tarafımdan bilimsel etik ilkelere uyularak yazıldığını ve yararlandığım eserleri kaynakçada gösterdiğimi beyan ederim.

22/ 06 / 2023

Aslı Islak İmzalıdır

Somayyeh SHABESTANI

☐ Tezle ilgili patent başvurusu yapılması / patent alma sürecinin devam etmesi sebebiyle Enstitü Yönetim Kurulunun ....../.../.... tarih ve ..... sayılı kararı ile teze erişim 2 (iki) yıl süreyle engellenmiştir.

☐ Enstitü Yönetim Kurulunun ....../.../.... tarih ve ..... sayılı kararı ile teze erişim 6 (altı) ay süreyle engellenmiştir.

## ACKNOWLEDGEMENTS

Firstly, I would like to thank to those who has been beside me throughout this challenging road.

Then, I am grateful to my supervisor Assoc. Prof. Dr. Merve GEÇİKLİ, thanks to whom this study could be completed.

I am also thankful to the members of my thesis committee Assis. Prof. Dr Zennure ELGÜN GÜNDÜZ and Assis. Prof. Dr. Veysel KARSLI, who gave very beneficial and valuable insights about my study. Besides, I would like to thank you Dr. Zeynep SÖNMEZ for professional help in statistical analysis

I would like to thank to all the participants who contributed to my study and provided very valuable data.

Somayyeh SHABESTANI

## ÖZ

### YÜKSEK LİSANS TEZİ

## MAKİNA ÖĞRENİM ALGORİTMALARI İLE İNGİLİZCE METİN ZORLUK SEVİYELERİNİN SINIFLANDIRILMASI: BİR YAZILIM ÖNERİSİ

Somayyeh SHABESTANI

Haziran 2023, 82 Sayfa

**Amaç:** Bu çalışmanın amacı, YDS okuduğunu anlama metinleri üzerinde alan uzmanları tarafından geleneksel formüllerle geliştirilen grid aracılığıyla yapılan metin sınıflandırma puanlaması ile önemli parametrelere dayanan modern yaklaşımlardan biri olan makine öğrenmesi algoritmaları kullanılarak yapılan metin sınıflandırma puanlamasını karşılaştırmaktır. Bu bağlamda çalışma, makine öğrenim algoritmalarının metin sınıflandırmasında kullanımının ölçme ve değerlendirme için metin okunabilirlik/zorluk düzeylerinin tespitinde kullanımına yönelik değerlendirilmesine dayanmaktadır.

**Yöntem:** Çalışma, alan uzmanları tarafından aynı örneklem (YDS okuduğunu anlama metinleri) üzerinden okunabilirlik formüllerine dayalı grid yoluyla yapılan metin sınıflandırma puanlaması ile modern yaklaşımlardan makine öğrenim algoritmaları kullanılarak yapılan metin sınıflandırma puanlamasının karşılaştırıldığı karşılaştırmaya dayalı nicel bir çalışmadır.

**Bulgular:** Gunning FOG Readability testi, Linsear Write, Flesch-Kincaid Grade Level, SMOG Index, Coleman-Liau, Dale-Chall-Readability, McAlpine EFLAW Readability ve Automated Readability Index olmak üzere sekiz okunabilirlik formülü kullanılarak YDS sınavlarından dokuz okuma-anlama metni okunabilirlik sınıflandırması için seçildi. Alan uzmanlarının kullandığı okunabilirlik formüllerine dayalı grid ile yapılan puanlama sonuçları, çalışmada kullanılan makine öğrenim algoritmaları temelli analiz yapan programın puanlama sonuçları ile istatistiksel olarak anlamlı farklılık göstermiştir. Ayrıca, makine öğrenimi algoritmasına dayalı program tarafından okunabilirlik sınıflandırmasının güvenilirlik puanı katsayısı oldukça yüksekken, alan uzmanlarının puanlamasının da güvenilir olmadığı da tespit edilmiştir.

**Sonuç:** Sonuçlar, makine öğrenmesi algoritmalarına dayalı program ile metinleri güvenilir bir şekilde sınıflandırırken, alan uzmanlarının metinlerin okunabilirlik sınıflandırmasına ilişkin verdiği cevaplarda daha fazla hata olduğunu göstermiştir. Bu nedenle, bu çalışmada kullanılan programın, yabancı dil eğitimi alanındaki paydaşları tarafından (İngilizce öğretmenleri, ölçme ve değerlendirme uzmanları vb.) dil öğrenenlerin seviyelerine göre okuduğunu anlama metinlerinin seçiminde kullanılması önerilebilir.

**Anahtar Kelimeler:** metin sınıflandırma, makine öğrenimi, YDS okuma-anlama metinleri

## ABSTRACT

### MASTER'S THESIS

#### CLASSIFICATION OF ENGLISH TEXTS 'DIFFICULTY LEVELS THROUGH MACHINE LEARNING ALGORITHMS: A PROGRAM RECOMMENDATION

Somayyeh SHABESTANI

June 2023, 82 Pages

**Purpose:** The aim of this study is to compare the text classification scoring made by field experts with the grid developed with traditional formulas on YDS reading-comprehension texts and the text classification scoring made using machine learning algorithms, which is one of the modern approaches over important parameters and with this scoring comparison. Accordingly, the study is based on to evaluate the use of machine learning algorithms in text classification according to the readability/difficulty level for measurement and evaluation.

**Method:** The study is a comparison based quantitative study in which the text classification scoring made by field experts through readability formulas based grid on the same sample (YDS reading-comprehension texts) was compared with the text classification scoring made using machine learning algorithms from modern approaches.

**Findings:** Using eight readability formulas: Gunning FOG Readability test, Linsear Write, Flesch-Kincaid Grade Level, SMOG Index, Coleman-Liau, Dale-Chall-Readability, McAlpine EFLAW Readability, and Automated Readability Index, nine reading comprehension texts were selected from YDS exams for their readability assessment. In the study, the scores of the readability formulas based grid used by field experts differ statistically significantly from the scores of the programme used in the study, which is based on machine learning algorithms. The reliability score of readability classification by the program based on machine learning algorithms is quite high while the scoring by the field experts is not reliable.

**Conclusion :** The results indicated that, with the programme based on machine learning algorithms classified the texts in a reliable manner while there were more errors in the answers provided by field experts regarding texts readability classification. Therefore, the programme used in this study might be recommended for stakeholders (English teachers, assessment and testing experts etc.) in ELT to use in the choice of the reading comprehension texts to the levels of language learners.

**Keywords:** text classification, machine learning, YDS reading-comprehension texts

## TABLE OF CONTENTS

|                                                   |      |
|---------------------------------------------------|------|
| KABUL VE ONAY TUTANAĞI.....                       | i    |
| ETİK VE BİLDİRİM SAYFASI.....                     | ii   |
| ACKNOWLEDGEMENTS .....                            | iii  |
| ÖZ.....                                           | iv   |
| ABSTRACT .....                                    | v    |
| TABLE OF CONTENTS .....                           | vi   |
| LIST OF TABLES .....                              | viii |
| LIST OF FIGURES .....                             | ix   |
| ABBREVIATIONS .....                               | x    |
| SECTION ONE .....                                 | 1    |
| Introduction .....                                | 1    |
| Related Studies and Rationale.....                | 1    |
| Aim and Research Questions of the Study.....      | 3    |
| Significance.....                                 | 4    |
| Transferability .....                             | 5    |
| Assumptions.....                                  | 6    |
| Limitations .....                                 | 6    |
| Delimitations .....                               | 7    |
| Definition of Terms.....                          | 7    |
| SECTION TWO .....                                 | 8    |
| Literature Review .....                           | 8    |
| Readability .....                                 | 8    |
| The Concept of Readability .....                  | 11   |
| Methods of Determining Readability.....           | 14   |
| Text Classification .....                         | 15   |
| Text Classification Applications .....            | 18   |
| Text Classification Methods .....                 | 19   |
| Text Classification Studies on English Texts..... | 30   |
| SECTION THREE .....                               | 32   |
| Methodology.....                                  | 32   |
| Research Method.....                              | 32   |

|                                                              |    |
|--------------------------------------------------------------|----|
| Data Source .....                                            | 32 |
| Population and Sample.....                                   | 32 |
| YDS Corpus .....                                             | 33 |
| Data Collection Techniques/Tools .....                       | 34 |
| Grid .....                                                   | 34 |
| Machine Learning Algorithms .....                            | 35 |
| Data Collection Process .....                                | 36 |
| Data Analysis .....                                          | 37 |
| Researcher Role .....                                        | 38 |
| Validity and Reliability .....                               | 38 |
| SECTION FOUR.....                                            | 40 |
| Results .....                                                | 40 |
| Findings Related to First Research Question .....            | 40 |
| Findings Related to Second Research Question .....           | 44 |
| Findings Related to Third Research Question .....            | 45 |
| SECTION FIVE .....                                           | 47 |
| Discussion, Implications and Conclusion.....                 | 47 |
| Discussion .....                                             | 47 |
| Implications.....                                            | 48 |
| Limitations and Suggestions for Further Research .....       | 49 |
| Conclusion .....                                             | 50 |
| REFERENCES .....                                             | 51 |
| APPENDICES .....                                             | 60 |
| APPENDIX-1. KATILIMCI RIZA FORMU VE BILGILENDİRME METNİ..... | 60 |
| APPENDIX-2. GRID.....                                        | 61 |
| APPENDIX-3. YDS CORPUS .....                                 | 63 |
| APPENDIX-4. ETİK KURUL RAPORU .....                          | 67 |
| CURRICULUM VITAE .....                                       | 70 |

## LIST OF TABLES

|                                                                                                                                                     |    |
|-----------------------------------------------------------------------------------------------------------------------------------------------------|----|
| <b>Table 1.</b> <i>Legibility, Intelligibility and Readability</i> .....                                                                            | 13 |
| <b>Table 2.</b> <i>T-test results of the difficulty classification scores of CATHOVEN and Experts for<br/>YDS Reading-Comprehension Texts</i> ..... | 41 |
| <b>Table 3.</b> <i>The Cronbach's Alpha of CATHOVEN Program</i> .....                                                                               | 45 |
| <b>Table 4.</b> <i>The Cronbach's Alpha of English Assessment-Evaluation Made by Experts Through<br/>Traditional Formulas</i> .....                 | 46 |



## LIST OF FIGURES

|                                                                                         |    |
|-----------------------------------------------------------------------------------------|----|
| <b>Figure 1.</b> <i>Document Classification With Artificial Intelligence</i> .....      | 16 |
| <b>Figure 2.</b> <i>Different Computer-Based Text Classification Methods</i> .....      | 20 |
| <b>Figure 3.</b> <i>The Difference Between Machine Learning and Deep Learning</i> ..... | 26 |
| <b>Figure 4.</b> <i>The Cathoven Software View</i> .....                                | 37 |



## ABBREVIATIONS

|               |                                                          |
|---------------|----------------------------------------------------------|
| <b>ARI</b>    | : Automated Readability Index                            |
| <b>BERT</b>   | : Bidirectional Encoder Representation from Transformers |
| <b>CCG</b>    | : Combinatorial Category Grammar                         |
| <b>C-Liau</b> | : Coleman-Liau                                           |
| <b>CNN</b>    | : Convolutional Neural Network                           |
| <b>F-KGL</b>  | : Flesch-Kincaid Grade Level                             |
| <b>KNN</b>    | : K-Nearest-Neighbor                                     |
| <b>LGBM</b>   | : Light Gradient Boosting Machine                        |
| <b>ML</b>     | : Machine Learning                                       |
| <b>NLP</b>    | : Natural Language Processing                            |
| <b>PCA</b>    | : Principle Component Analysis                           |
| <b>SMOG</b>   | : Simple Measure of Gobbledygook                         |
| <b>SVM</b>    | : Support Vector Machine                                 |

## **SECTION ONE**

### **Introduction**

The focus of the current research is to provide a solution to eliminate the person-oriented scoring bias in the classification and to minimize the margin of error with valid and reliable measurements of the readability levels of English texts and in vectorization. In this condition, the main contribution to foreign language education in general and foreign language education in Turkey, in particular, is to present its advantageous aspects by adopting the most widely used data-based approach in order to make use of a program, which is becoming increasingly widespread in the universal context in foreign language education, measurement-evaluation in the context of Turkey and to encourage assessment and evaluation experts as well as language educators to use this program. On this basis, in this section, the related studies and rationale of the current study are mentioned. Then, the details about the aim and the research questions guiding the study are given and the significance of the study is discussed. The section continues with the discussion of limitations, delimitations, and assumptions. The introduction part concludes with the definition of the key terms.

### **Related Studies and Rationale**

Due to its unstructured nature, text is an extremely rich source of information, but extracting insights from it can be challenging and time-consuming. A text classification or categorization involves organizing open-ended text logically and assigning predefined categories to it. Many fields, such as document retrieval, routing, and clustering, use it to organize, structure, and categorize texts. Classification of English texts with data-driven smart program is one of the applications of machine learning algorithms that provide data about English texts as a result of input-output processing in natural language processing (NLP). In this context, data-based machine learning algorithms determine the readability/difficulty levels of texts and make classifications by analyzing the elements specific to the nature of the language in natural language processing, thus drawing a map of where and how a text will be used (Miltakaki & Troutt, 2007). It is clear that in the digital age, there is a wealth of digital resources as well as print resources for English language teaching and assessment; The main thing here is to choose the appropriate text for the level, whether it is measurement and evaluation or course teaching material in the foreign language education-teaching process (Ortego & Sanchez, 2019). The key dimension in classifying whether the texts are suitable for

the level is the analysis of the readability of the text, textual classification of texts to the certain levels (Huerta, 1958). Studies in the international literature show that in this context, smart data-based soft wares following certain formulas, which has become increasingly common in text classification for the last 10 years, has begun to replace expert-oriented traditional methods in the classification of English texts for various purposes in different disciplines, including education (Hancke et al., 2012; Hirschfeld & Gelman, 1994; Kanungo & Orr, 2009; McConnell, 1982). From this point of view, this thesis proposal is structured to encourage the selection of traditional formulas based algorithms and to explain the functionality of this mechanism in order to determine the appropriateness of the texts to the level in language education in the context of Turkey where English is taught as a foreign language. Focusing on readability/difficulty levels is presented as the main criterion in classifying English texts according to Search metrics (Graesser et al., 2004; McNamara et al., 2002; McNamara et al., 2010). In this context, as seen in the literature, it covers the fields of Quantitative and Statistical Algorithms and Informatics, which include examining and evaluating the legibility/difficulty level and examining the quantitative effects of knowledge. With both a practical and theoretical scope, these scientific disciplines, in addition to emphasizing the development of mathematical models, pay particular attention to the derivation of measurements for the different phenomena studied, namely the semantic, syntactic, lexical, and topological of the language, and the necessity of being inclusive. It was noted by Luo (2021) that the size of the dataset affects the advantages and disadvantages of machine learning classification. According to Bhavani & Kumar (2021), determining the appropriate structure and technique for text classification is challenging. An experiment by Liu & Wang (2020) demonstrated that neural text classification based on quality classification is 96% accurate. As Lv & Liu (2005) pointed out, semantic mining plays an important role in text classification. In addition to helping scientists identify texts' levels and analyze them, text classification helps them organize according to their understanding level.

Although traditional formulas offer many models and grids in such measurements, and the grid diversity is very rich, the high margin of error in the measurements and the fact that the classifications do not reflect the fundamental nature of the text showed that there is a need for more advanced systems with intelligent systems (Hancke et al., 2012; Kanungo & Orr, 2009). Another issue related to traditional formulas is that they need time to be implemented as a methodology, require more than one or even more raters for valid-reliable measurement to ensure validity-reliability between raters, and sometimes the data become accessible only to experts due to the complexity of the equation used.

Since advanced machine algorithms operate intelligent measurement systems and versatile statistical analysis techniques on the text simultaneously, the time spent in text analysis and classification has decreased significantly. The textual data mining provided by these algorithms determines the text's real readability/difficulty level by making each parameter's difficulty level on a partial and holistic basis. Since it is based on precise measurements and is independent of the person, the margin of error is very low. In addition, since it can analyze very comprehensive textual documents, it provides valid and reliable analyses in both narrow and large corpora. Another advantage is that it allows comparison between corpus and measurements (Tseng et al., 2019; Moreno & Redondo, 2016; Vijayarani et al., 2015).

Accordingly, since it is recommended in the literature to use the digitally ready corpus for the inputs ready for the coding of the program system in natural language processing (Maqsood et al., 2020), the study was carried out on a corpus of a high stake exam's reading comprehension questions in Turkey to determine the level of textual readability. So, it is based on the focus on selecting texts suitable for the level in assessment and evaluation. In this context, it is planned to compare the classifications of English assessment experts and a program based on traditional formulas oriented algorithms based and to discuss the role of machine learning algorithms at this point.

### **Aim and Research Questions of the Study**

In this study, it is aimed to compare the text classification scoring made by field experts with the grid developed with traditional formulas on YDS reading-comprehension texts and the text classification scoring made using a program based on traditional formulas algorithms, which is one of the most frequently used mechanism and with this scoring comparison, it is aimed to evaluate the use of traditional formulas in text classification according to the readability/difficulty level for measurement and evaluation. In this context, data mining provided by traditional formulas based program through text classifications made by tokens based on syntactic, lexical, semantic, and topological parameters, and the information provided by data mining in determining the readability/difficulty level were studied (Ortego & Sanchez, 2019). For this purpose, the main research question that guided the study is as follows:

1. Is there a significant difference between the difficulty classification of YDS reading-comprehension texts with a program using the modern data-based methods (machine learning algorithms) and the classification made over traditional formulas by field experts?

2. What is the reliability score of the difficulty classification scoring of YDS reading-comprehension texts in the measurement made with a program using modern data-based methods (machine learning algorithms)?
3. What is the reliability score of the difficulty classification scoring of YDS reading-comprehension texts in the measurements made by English assessment-evaluation experts through traditional formulas?

### ***Significance***

In foreign language education and language studies, readability, in other words, the difficulty or complexity of English texts or text groups (Macahilig, 2014) plays a key role in reading comprehension, both in the selection of appropriate assessment and evaluation texts for the level and they are among the factors that affect the selection of appropriate course materials. As a matter of fact, many studies in the related literature show that there is a positive and significant relationship between the readability of a text in a foreign language and reading comprehension (Dale & Chall, 1949; Guthrie, 2007; Kintsch & Vispond, 2014; Routman, 2003; Snow, 2002; Stevens et al., 2017). This makes it necessary to carry out text classification studies in selecting reading-comprehension texts and text groups, taking into account the level of legibility in measurement and evaluation in a foreign language at the education-teaching level. Because text classification studies give the opportunity to categorize a text multidimensional according to difficulty/complexity levels in terms of readability with predefined criteria and parameters (Joseph & Ramakrishnan, 2015). However, at this point, educators have problems with how to classify and seek solutions in terms of which legibility classification method provides valid and reliable data.

In this context, the solution proposals presented in the literature are presented through two basic methods; the first of these is the classification of the texts by scoring the educators with the grids structured with traditional formulas, the second one is intelligent program-assisted classification using machine learning algorithms, which is one of the modern approaches with data (Danielson, 1987; Maqsood et al., 2020; Roman et al., 2021; Rush, 1985). Grids based on traditional formulas, especially those in the first generation (Danielson, 1987), are frequently preferred by educators in text classification because they contain computationally simple mechanisms (Rush, 1985). However, when traditional formulas are examined as the parameters they measure, they have a mechanism without content validity in terms of difficulty/complexity level in English; In other words, since traditional grids focus on surface elements (such as syllable number, word number, sentence length), metacognitive (e.g. metadiscoursal elements) elements, which are other determinant dimensions of complexity in

texts, are not included in the scope, which means that both measurement and measurement in scoring and classification are not included in the scope. It has been determined that the tool causes erroneous measurements and classifications due to the scorer and the error margin (Maqsood et al., 2020; Roman et al., 2021). Although the studies on traditional formulas are progressing towards eliminating the deficiencies and increasing the content validity, they were insufficient to present valid and reliable data about the readability of the text in classification since micro and macro structures could not be adequately parameterized (Si & Callan, 2001).

In this context, the formulas that have been improved closest to natural language processing in classification based on traditional formulas are the grids developed as a result of third-generation studies and include a more holistic inclusive structure (François & Fairon, 2012). However, it has been observed that as the text group expands as a corpus, in which these grids are more effective in narrow corpora, which are small-scale data sets, errors occur in classification, and the classification criteria are insufficient in scope (Graesser et al., 1994). These problems experienced in measurements based on traditional formulas in text classification have prompted experts to develop more advanced mechanisms, and together with the developments in information and communication technologies, machine learning algorithms, known as data-based intelligent program systems, and called the era of modern approaches in text classification, have emerged (Li et al., 2018).; Rosyid, 2018; Vijayan et al., 2017). The main starting point of this study is to give the working principles of machine learning algorithms and to test the scoring efficiency, in contrast to the traditional formulas that cause the educators to choose the text that is not suitable for the level in determining the readability with faulty or missing measurements in English text classification.

### ***Transferability***

It is the purpose of this study to evaluate how machine learning algorithms can be used to classify texts according to readability/difficulty levels for measurement evaluation, by comparing text classification scores made using traditional formulas and machine learning algorithms on a given parameter basis. Both at the educational level and at the assessment-evaluation level in foreign language instruction, a text-based data mining effort will be conducted. The results achieved in the design and selection of appropriate measurement tools and training materials will serve as a guide for practitioners, educators, program and curriculum development experts in the field. Additionally, it will be determined which aspects of the syntactic, semantic, topological, and semantic dimensions of English need to be improved in the selection of texts appropriate for the level based on whether or not the expert classifications based on conventional formulas are accurate and impartial. Decision-makers and practitioners

will be able to reassess their options and choose strategies and methods that are suitable for contextual English education at the level and demands of A1-A2/B1-B2/C1-C2 by contrasting text classification options using machine learning algorithms.

As a result, the project will make a pedagogical, valid-reliable, and time-saving contribution to the context-based English education methods in Turkey, which are essential to fulfilling the European Language Portfolio objectives. In this study, we will present a science-based, systematic, and intelligent systems-based guide on selecting texts. Additionally, machine learning algorithms will pave the way for its application in various research in order to generate suitable instructional materials and measurement instruments for measurement-evaluation related to the teaching of English as a foreign language, and good advancements will be seen.

### ***Assumptions***

Participants were assumed to attend the study voluntarily in order to code the texts on the basis of readability level. Besides, it was also assumed that they objectively coded the texts without any bias and all of them had Internet access to participate in the study. Another assumption is that the academicians, who participated, were testing experts. Moreover, it was assumed that the academicians would allow the researcher to use the data they provided by signing consent forms before the study. Furthermore, it was assumed that the program used in the study would classify the texts' readability reliably and validly.

### ***Limitations***

The primary limitation of this study was the voluntary nature of the online response, which unfortunately the participation in the online system was less despite many follow-ups and sending reminder emails to academicians. Of course, the reason for this could be the coincidence of the schedule with the new year and also the busy schedule of the academicians could result in fewer participation.

Another limitation was that not all of the academicians who participated in the study were at the same academic level; some were assistant academicians and others were full academicians. In addition, some of them were graduates of native English-speaking universities, which could have caused inconsistent results. The reader's familiarity with the reading passage and interest in the text's subject matter and content may actually be more crucial than the text's degree of lexical and grammatical complexity. As our target community are faculty members specializing in ELT, they will most probably find most passages easy enough, and this will overshadow our research findings.

The other limitation of the study was the limited number of texts focused. For more generalizable findings, the corpus to be compiled should be larger. However, the textual length was considered in order to motivate the academicians to code the texts, which could have caused a cautious interpretation of the findings.

### ***Delimitations***

In this study, only the academicians, expert in testing, participated in order for scientific, reliable, and valid coding of textual readability and classification because testing experts have relevant background, knowledge, and experience in the identification and analysis of reading texts and to identify the reliable texts to the level of students. Therefore, their expertise should be considered so that to compare and contrast between human-based coding and machine coding could be done. In addition, it was important to investigate if there would be any variation human based coding and machine coding. Moreover, it was essential to explore, if there is, any possible significant difference among codings to measure the reliability of two different readability scores. Finally, it was more appropriate to focus on YDS texts as high stake exam in Türkiye because it is significant to identify reliable readable texts to measure reading skill of the participants.

### **Definition of Terms**

#### *Readability*

As explained by the Oxford Dictionary (2023), Readability (/ˌriːdəˈbɪləti/) is the fact of being easy, interesting and enjoyable to read.

#### *Classification*

According to the Oxford Dictionary (2023), Classification (/ˌklæsɪfɪˈkeɪʃn/) involves the act or process of putting people or things into a group or class (= of classifying them).

#### *Machine Learning*

As described by the Oxford Dictionary (2023), Machine Learning (/məˈʃiːn ˈlɜːnɪŋ/) involves a type of artificial intelligence in which computers use huge amounts of data to learn how to do tasks rather than being programmed to do them.

## SECTION TWO

### Literature Review

#### Readability

The readability/difficulty assessment of the text is an important practice in text mining (A text mining process, also known as text analytics or text data mining, is the process of extracting high-quality information from text. This is the process of discovering new, previously unknown information through the automatic extraction of information from a wide range of written resources. It is an artificial intelligence technology that combines natural language processing (NLP) and machine learning (ML) to transform free (unstructured) text into normalized, structured data), which can be considered more formally as the sum of all the factors that affect the reader's comprehension, reading speed, and level of interest in the textual material (Dale & Chall, 1949). Effective assessment of the difficulty of texts can contribute to teaching and learning (in Education and Training): readers will understand and learn effectively when they choose texts that are suitable for their reading abilities (at their level) (Klare, 1963, 2000). Texts that are too advanced for one's reading level (texts that are well above one's level of readability) will significantly increase the reader's cognitive load and cause feelings of frustration (DuBay, 2004). On the contrary, texts that are far below one's level in terms of readability may lead to a lack of motivation and a sense of accomplishment in the reader. In other words, the readability assessment helps to match reading skills with reading materials from which more successful and enjoyable reading experiences can be created. In the text readability/difficulty classification, two methods are followed in the literature: traditional and data-driven methods (Machine learning-based). Traditional readability formulas are based on research findings that factors such as semantic, syntactic, and lexical difficulty affect the difficulty level of a text (Graesser et al., 2011; Rubenstein & Aborn 1958). Therefore, the linguistic properties of the readability formulas are used as key variables to predict the readability of the text. For example, the Flesch Easy Reading Formula (Flesch, 1948) accepted the number of syllables in a word as an indicator of lexical complexity and sentence length as an indicator of syntactic complexity and measured text difficulty based on the average number of syllables per word and average sentence length (Collins-Thompson, 2014).). The more syllables an average text word has and the longer the sentence, the harder the text. Chall & Dale (1995) added the percentage of difficult words as another variable of text difficulty – the more difficult words in a text, the harder it is. As Graesser et al. (1994) stated, traditional formulas

that use general language features fail to reflect the actual reading process. This is because while it contains general language features such as meaning, syntax and lexical features, it neglects other basic text features. In addition, Collins Thompson (2014) argued that due to traditional readability formulas, the deep content features of the text are neglected and the focus is on shallow information. Using such conventional readability formulas leads to skepticism in predicting the intelligibility of the text. For example, in order to decide whether a word is difficult in a text, it is not sufficient to determine the frequency of the words in the text, since the use of words develops over time and new words emerge. For this reason, the Thorndike List developed by Thorndike & Lorge (1944) based on word frequencies is not ideal for assessing the difficulty of modern words. Therefore, it makes the predictions of readability formulas less accurate. Word frequency is not an absolutely efficient way to determine word difficulty. For example, the word toothbrush is simple, widely used and understandable. However, this phrase is classified as a difficult word since it is rarely seen in written texts, newspapers, magazines or textbooks. However, this is absolutely not true. Using word length to calculate readability has been questioned because longer words are not always more difficult (Yan et al., 2006). Basing syntactic complexity on sentence length and using it as a variable to measure readability has been criticized as being too intuitive and illogical (Bailin & Grafstein 2001). Scientists have conducted research using various language features. For example, Graesser et al. (2004) developed Coh-Metrix, an online analyzer of text features using categories such as vocabulary, syntax structure, implicit semantics, and cohesion. The effectiveness of Coh-Metrix readability features for text classification is supported by empirical evidence in psychology (Graesser et al., 2004; McNamara et al., 2002; McNamara et al., 2010). On the other hand, Kanebrant et al. (2015) presented T-MASTER, a tool to assess student's reading skills in various dimensions. They used the SVIT model to analyze textual complexity in T-MASTER. This model included four categories of variables, namely surface complexity, vocabulary difficulty, syntactic and idea density, and was created. As a result of further developments in the index in data-driven methods (machine learning-based) in recent years, the program has been developed for valid and reliable definitions of sentence complexity, taking into account additional complex features such as word frequency, lexical meaning, text morphology, and the depth of the parse tree. The system is based on SVM (Support Vector Machine) to calculate the complexity of the sentence. This system considers lexical, syntax and morpheme features to categorize text. With the rise of natural language processing techniques and machine learning algorithms, it allows researchers to refine model algorithms to measure text readability with a more flexible scope (Feng et al., 2010; François & Miltsakaki, 2012; Petersen & Ostendorf, 2009; Sung et al., 2015; Vajjala & Meurers, 2012). In addition to the studies mentioned above on document readability,

some scientists have begun to examine readability at the sentence level. For example, Vajjala & Meurers (2014) used 151 features to train readability models at both the document and sentence level, built on their previous research (Vajjala & Meurers 2012). Experimental results show that the document-level readability model achieves a Pearson correlation of 0.92, but the sentence-level readability model only achieves an accuracy of 66%. Also, Ambati et al. (2016) used an incremental Combinatorial Category Grammar (CCG) parser to calculate sentence complexity and estimate relative sentence readability. The incremental CCG model outperformed the non-incremental Sentence Structure Trees (PST) model in extracting syntactic features. On the other hand, Howcroft & Demberg (2017) used the Model Blocks parser and icy-parses program to extract 22 features. Four measures (idea density, surprise, cost of integration, and depth of insertion) were used to calculate sentence complexity scores, creating an effective and efficient basis for training different readability models to find out which feature set could constitute a readability model. The analysis results show that the depth of surprise and embedding makes the readability model more efficient when compared to integration cost and idea density. For further reading, Collins-Thompson (2014) may be consulted for a comprehensive overview of related studies that follow this line of research.

The ever-increasing growth of databases in every area of human activity has increased the need for powerful new tools to transform data into useful knowledge. To meet this need, researchers have explored methods and ideas in various areas such as machine learning, pattern recognition, statistical data analysis, data visualization, neural networks, econometrics, information retrieval, information extraction, etc. For this purpose, the classification of Latin texts based on a support vector machine is used in this thesis. Categorizing or classifying texts means assigning a text document to a predetermined appropriate class. The main challenge of text classification is the size of the feature space in such problems. In many existing algorithms, such a large space will lead to a lot of slowness of the cluster and its inefficiency. In addition, some features not only cause better categorization of texts but also reduce the accuracy of categorization. Various methods have been presented in connection with feature selection techniques and learning algorithms (Kou et al., 2020; Kowsari et al., 2019; Larose & Larose, 2014; Reddy, 2011; Mayfield & Penstein-Rosé, 2010).

For this purpose, first remove the stopping words and then root the words and weight the features, and also PCA method was used to reduce the features. Today, with the expansion of database systems and the large volume of data stored in these systems, there is a need for a tool to process the stored data and provide the information resulting from this processing to the users. By using simple queries in 1SQL and various common reporting tools, information can

be provided to users so that they can conclude the data and the logical relationships between them. However, when the volume of data is high, even though the users are powerful and experienced, they cannot recognize useful patterns among the large volume of data, or if they are able to do so, the cost of operations in terms of human and material is very high.

### ***The Concept of Readability***

Readability is a language-specific concept. In the ninth century BC, the clergy employed it for the first time to categorize the qualities of the words that make up sacred scriptures, taking into account the quantitative characteristics of the words (Mirzaolu & Akn, 2015). Later, inter-institutional texts, health and military organizations, and the sphere of education began to adopt the notion of readability (Boztaş et al., 2014).

The literature contains numerous definitions of readability. According to Ateşman (1997), readability refers to how easy or difficult it is for the reader to understand the content. In other words, a text of a reader with a certain level of education; It is the situation in which the reader can read, understand or the language-specific features in the text can be made meaningful by the reader (Güyer et al., 2009). Another definition of readability states that it is the essence of the text's clarity and the simplicity with which the words and phrases may be understood (Tekşan, 2016).

Readability is a concept related to the difficulty of the texts and the level of compliance with the level of the reader (Güyer et al., 2009). The word and sentence lengths are typically utilized as two criteria to determine a text's level of complexity and ease. In this sense, readability gives a notion about the intelligibility of the text by a particular level of reader by employing the features of syllables, words, and sentences of a language (Tekşan, 2016). In general, multiple and average number of syllables, word frequency, average word length and average sentence length can be counted as factors affecting readability. Increasing the number of words in a sentence reduces the readability of the sentence by the reader (Boztaş et al., 2014). According to Gunning (1952), readability is a useful tool for figuring out what makes a document difficult to read. Dubay (2004) states that readability makes some texts easier to read and this concept is sometimes mixed up with legibility. Richards & Schmidt (2010) defined readability as a measure of how easily the text can be read and understood, and focused on the variables that reflect the average sentence length, the number of words in the sentences and how complex the language used is in readability.

A text's readability is influenced by both the reader's traits and the content of the text. These characteristics include the reader's language skills, subject-matter knowledge, and

interest in the content (Al-Tamimi et al., 2014). One of the key components of literacy is the reader's reading proficiency or educational level (Badarudeen & Sabharwal, 2010).

Literacy levels can be divided into 4 as follows:

- Illiterate: People who cannot read very simple words and paragraphs.
- Not fully developed (basic, beginning): A person who cannot find tacit knowledge conveyed in a simple and familiar text.
- Basic: A person who is functionally literate. For example, a person who can read and understand a text of average length and find information that can be determined by inference.
- Advanced: Someone who is fully literate. For example, it is a person who can read long texts, the parts that these texts are related to, and make inferences and syntheses (Aluisio et al., 2010).

Today, search engines and other web tools include readability algorithms, the most used technique for determining the readability level. As a result, readability levels were associated with web results, and better general writing quality was encouraged. In addition, formulas measuring the level of readability; A better readability score is aimed by adding it to most text editors such as Microsoft Word, KWord, IBM Lotus Symphony, WordPerfect and WordPro (Al-Tamimi et al., 2014).

For example, the following steps should be followed to evaluate the readability of a text using the Flesch Reading Ease and Flesch-Kincaid Education Levels readability formulas in Microsoft Word:

- The digital copy of the written material must be saved as a Word document.
- An article on a web page should be saved as plain text.
- All off-topic information such as dates, copyright sentences, postal addresses should be deleted.
- Only the text itself should remain, and headings, subheadings, tables and figures should be deleted.
- “colon”, “semicolon” and decimal points in numbers should be deleted in a sentence.
- “Options” must be selected under the “tools” menu in the Microsoft Word Office interface.

- Click on “Spelling and grammar” and check the “Show readability statistics” box.
- In this way, spelling and grammar check is done on the file. When the text is complete, a window opens showing the text's Flesch Ease of Reading and Flesch-Kincaid scores (Badarudeen & Sabharwal, 2010).

Although readability is one of the variables regarding how suitable a text is for its reader, it is the most important one (Freda, 2005). In other words, readability is a problem related to the harmony between the text and the reader. While simple texts are boring for a strong reader; On the other hand, texts that cannot be read fluently are stopped from being read by a weak reader. Also a reader; have difficulty in reading poorly printed texts that contain complex and long sentence structures (Solmaz, 2009).

Legibility and intelligibility, two terms that sometimes get mixed up in the literature, are strongly related to readability. Legibility includes only the external structure features of a text (Okur & Ari, 2013). In other words, formal elements including color, font size, and positioning of tables and figures inside texts all contribute to legibility. The semantic characteristics of a document have an impact on its understandability. The quantity of syllables or phrases, challenging or foreign words, technical terminology, and abstract and tangible words come first among these characteristics (Güyer et al., 2009). The difference between readability and intelligibility can be given as follows: Readability is the characteristic of a text itself while intelligibility is an indication of the reader's ability to make a text meaningful.

In addition to the syntax, the reader's background knowledge, interests and general reading level are also effective in order for a text to be understandable while it can be read (Wissing et al., 2016). Table 1. shows the concepts of legibility, intelligibility and readability as a whole.

**Table 1.** *Legibility, Intelligibility and Readability*

|                 |                                                                                                                              |
|-----------------|------------------------------------------------------------------------------------------------------------------------------|
| Legibility      | External structure features (font size, color, table and shape, composition of the page, etc.)                               |
| Intelligibility | The content of the text and its internal structure features (text subject, plan, language, expression, etc.)                 |
| Readability     | Determining the readability and reading difficulty level of the text by looking at the syllable, word and sentence structure |

The texts' qualitative and quantitative characteristics should be highlighted throughout analysis. A reader's willingness to continue reading a text indicates that he or she understands the text. However, a text that is understandable by the reader does not mean that the text is

readable (Güyer et al., 2009). In this view, the factors that affect how well a reader comprehends a text include error-free content, an engaging layout, active language, images, enough white space, bulleted lists, and cultural adherence (Freda, 2005).

Badarudeen & Sabharwal (2010) suggest the following ways to improve the readability and understandability of patient education materials:

- The number of messages to be given in the material should be reduced.
- Short sentences should be used.
- Language should be used as if you are talking to someone.
- The use of medical jargon should be reduced.
- Ambiguous words should be avoided.
- Common words should be chosen and used consistently
- Common and culturally appropriate analogies for the target reader should be used.
- When giving statistics on the subject, words such as “half”, “one-third” should be used instead of real numbers.
- Quotation marks and symbols should be avoided (Badarudeen & Sabharwal, 2010).

### ***Methods of Determining Readability***

To determine the readability level of texts, three methods are commonly used: expert opinion, subtractive readability method and readability formulas. Expert opinion is based on taking the opinion of an expert on the text's readability level. A disadvantage of this method is that it takes a long time to determine readability and is subjective (Karatay et al., 2013). Among the readability formulas, sticker readability is the most widely used method. In spite of the readability formula's ability to measure the difficulty of a written text, it is insufficient to measure its intelligibility.

This is due to the fact that these formulas are based on sentence lengths and word lengths. In this logic, long sentences are harder to understand than short ones (Fanguy et al., 2004). Words or sentences that are simple; It fails to capture the precision of word choice to accurately express measurable phenomena such as the limitations of common words, the tendency to use certain word forms, the significance of combining tools, or the rate at which economic variables change (Greene, 2001).

To address these shortcomings, Taylor (1953) devised the subtractive readability approach to assess the readability of texts published in the mother tongue (Keshavarz & Salimi,

2007). Fill-in-the-blank tests for discrete sentences are different from sticker readability tests. The subtractive readability method considers both the subject and the surrounding factors when applied to lengthy texts (Keshavarz & Salimi, 2007). The sticker readability method is simple and easy to prepare (Bormuth, 1968). The literature also mentions the disadvantages of the subtractive readability method, such as how time-consuming and difficult it is for readers, and how few texts can be examined on each reader (Kandula & Zeng-Treitle, 2008).

## **Text Classification**

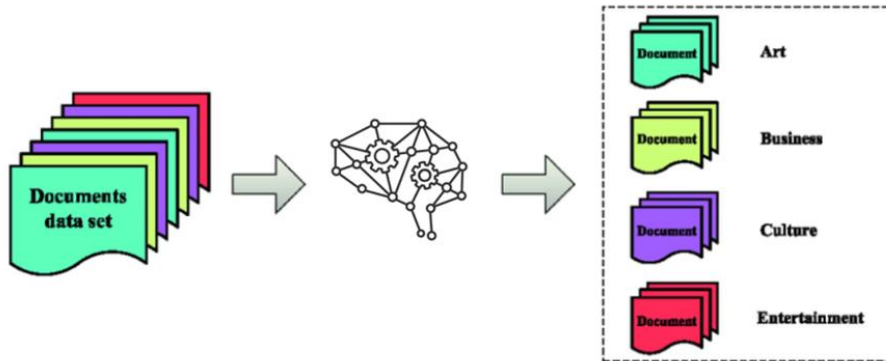
The need for strong new tools to turn data into knowledge is growing as databases in every aspect of human activity continue to expand. Researchers have investigated techniques and concepts in a variety of fields, including machine learning, pattern recognition, statistical data analysis, data visualization, neural networks, econometrics, information retrieval, information extraction, etc., to address this need. One of the "Intradisciplinary" difficulties in the fields of library science, computer science, and information science is text classification, often known as "text tagging" or "text categorization" (Kavitha and Prabhavathy, 2021). It is one of the challenging complex systems, which is able to classify new documents into predefined classes using several techniques and their combinations (Mirończuk and Protasiewicz, 2018). Moreover, to achieve higher computational efficiency, researchers trying to develop new classification systems or improving the existing ones (Pinheiro et al., 2017). Categorizing or classifying texts means assigning a text document to a predetermined appropriate class. A schematic of document classification using artificial intelligence is shown in Figure 1.

A machine technique called text categorization allocates the desired text to a collection of predefined sets that the core of many program systems are text classification algorithms. It is estimated that about 80% of all content on the web is unstructured data. Since text is one of the most common types of unstructured data, organizing, categorizing, and analyzing text data is a challenging and time-consuming operation. Natural language processing and machine learning techniques come to the aid of companies and program manufacturers to do such tasks faster and more cost-effectively (Debortoli et al., 2016; Kumbhar & Mali, 2016; Pop, 2016). And aimed to systematically classify or categorize text data into specific and predetermined groups (Wu et al., 2014; Wahdan et al., 2020).

Document classification is one of the prevalent issues in computer science and libraries nowadays. During the last 20 years, digital text documents have grown exponentially that one of the results of this exponential growth, is the significance of document classification based

on content. One of the analytical methods used in artificial intelligence is document classification.

**Figure 1.** *Document Classification With Artificial Intelligence*



The vastness of the feature space in these situations is the primary obstacle to document categorization. Such a large space will result in a significant slowing of the cluster and its inefficiency in many of the existing methods. Additionally, some factors not only improve document classification but also decrease classification accuracy. A text must be indexed in order for a classifier or classifier algorithm to understand it; otherwise, the text cannot be understood (which expresses its contents in dimensions). This crucial aids in giving the texts of the educational, experimental, and validation sets the uniformity and essential form (Zhang et al., 2021; Abualigah et al., 2021; Kowsari et al., 2019).

In connection with feature selection techniques and learning algorithms, various methods have been presented. Information has a lot of value to us, with the increase in the amount of information available on the Internet, there is a great need for tools that can help in searching, filtering and managing resources. Given the huge volume of electronic textual information that is significantly available through the Internet and other sources, the work of collecting and processing uncategorized textual information faces many challenges in the lack of effective indexing and classification (Cai et al., 2018; Vijayarani et al., 2015).

There are various purposes for classifying texts, including document management, document enlargement, and information volume reduction. Text classification is the process of categorizing natural language documents according to a predetermined set of themes. And it is used in a variety of applications, including text filtering, automatic metadata generation, word disambiguation, the creation of hierarchical catalogs from web resources, automatic systems for answering questions, determining the relevance of data, and electronic mails. Text classification is also used in many other fields. Other relevant topics include appraisal and title recognition. Additionally, it is employed in any application that calls for the organization of

papers or a unique, adaptive distribution of the aforementioned materials (Rasane & Patil, 2016; Rostami et al., 2015).

As it has been mentioned, database systems have become more prevalent in businesses today and have led to the storage of a significant amount of data. Due to this incident, there are now more demands for data processing and a platform to deliver the processes' outcomes to consumers. When the amount of information increases, users cannot easily discover the patterns between the data without appropriate tools, if they want to achieve this, they must invest a significant amount of money and human resources. To address these issues, there is a need for systems that can discover knowledge with minimal user intervention and automatically express logical patterns and relationships (McGilvray, 2021; Al Nuaimi et al., 2015).

Information has much value to us; with the increase in the amount of information available on the Internet, there is a great need for tools that can help in searching, filtering, and managing resources. Considering the vast amount of electronic textual information that is significantly accessible through the Internet and other sources, in the absence of proper indexing and classification, the task of retrieving and processing uncategorized textual information faces many problems. Classification of texts has many uses, including document tracking, document management, document expansion, and information volume reduction. Text classification refers to the act of thematic labeling of natural language texts based on a predetermined set. From the applications of text classification in many fields, from text indexing based on a controlled dictionary to text filtering, automatic metadata generation, word disambiguation, generation of hierarchical catalogs from web resources, automatic systems for answering questions, recognizing the relevance of data, electronic mails. Appraisal, title recognition and other related fields can be mentioned. And in general, it is used in any application that requires the organization of documents or a special selective and adaptive distribution of the documents in question. The main challenge of document classification is the feature space size in such problems (Devi & Kumar, 2020; Kowsari et al., 2019; Rousseau et al., 2015; Franks, 2013).

In many existing algorithms, such a large space will lead to a huge slowdown of the cluster and its inefficiency. In addition, there are features that not only cause better classification of documents but also reduce the classification accuracy. A text cannot be interpreted directly by a classifier or a classifier algorithm but by using an indexing process that maps the text to an index (which expresses its contents in dimensions). This important help to provide uniformity and a necessary form for the texts of the educational, experimental, and validation set. The large dimensions of the term space in the category of texts are usually

troublesome. In fact, as the term space increases, the number of features also increases, which on the one hand, causes more complexity (spending more time and memory space), and on the other hand, the lack of dependence between information decreases. Data that are dependent on each other are generally not worth categorizing. In other words, in reducing the general dimensions, an attempt is made to determine the terms that have little value in the application by analyzing the linguistic structure and all the texts in the educational collection, and these terms are determined in the form of a fixed list. This list automatically filters the terms of the input text. The same work is done for each category in a local area separately. Therefore, the problem of reducing dimensions is one of the good fields in information retrieval and especially text classification. The first approach for dimensionality reduction using term selection is called the filtering approach. Using the tools provided by statistics or information theory, irrelevant terms are filtered from the extracted terms. Finally, the categories are produced independently of the filter function used, using the reduced term space. Due to the vast amount of electronic textual information that is significantly accessible through the Internet and other sources, in the absence of proper indexing and categorization, the task of retrieving and processing uncategorized textual information faces many problems. Many machine learning methods have been used in the field of text classification in recent years, including regression models, k-nearest neighbor classification, Bayesian networks, support vector machine, decision tree and inference learning techniques, each of these methods have different accuracy and calculations. In the combined methods of text classification that include several individual categories, the goal is to use several categories instead of having one category. Studies show that by using the combined classification methods, the quality of classification can be improved, while with individual methods, the quality of classification cannot be increased beyond a certain limit. In this regard, a new combined method for classifying texts is presented in this thesis.

### ***Text Classification Applications***

#### **I. Indexing for Boolean information retrieval systems**

Most of the common applications in this field of research for automatic indexing of texts for information retrieval systems are based on a controlled dictionary, the prominent examples of which are Boolean systems. For each text, one or more keywords or key phrases that describe its content are determined.

#### **II. Organization of texts**

Indexing with a controlled dictionary is an example of the more general problem of text-based organization. In general, many discussions related to text organization

and archiving for personal, organizational, or text-based legal structures have been conducted through text classification techniques.

### III. Filtering texts

Text filtering is one of the classification activities, which is a string of input texts distributed by an information producer to an information consumer asynchronously. Filtering can be seen in the form of a category of single-label texts, in which categories of entered texts are divided into two separate categories (related and unrelated).

### IV. Disambiguation of the word

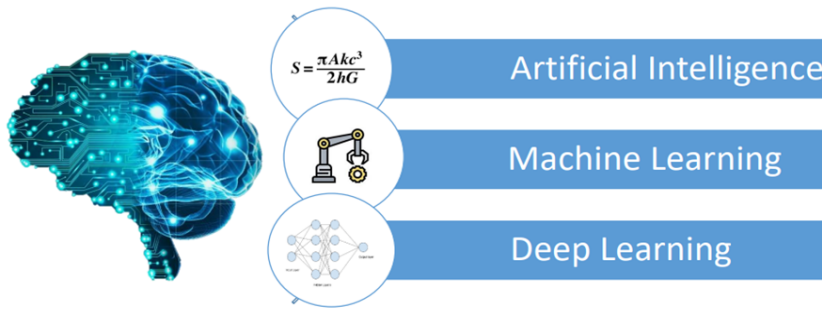
Eliminating the word's ambiguity is said to find the correct cognate words in a text. Disambiguation is important for many applications, such as natural language processing and document indexing, using the preference of the role of the word over the word itself in the field of information retrieval.

## ***Text Classification Methods***

In order to anticipate categories of text data based on these qualities, text classification seeks to extract features from raw text data. Over the past few decades, numerous models have been proposed for text classification. NB is the initial model utilized in traditional models for the text classification task. After this, general classification models such as K-nearest-neighbor (KNN) (Boateng et al., 2020), SVM (Assegie, 2021), and Random Forest (RF) (Dong et al., 2019), which are called classifiers and used broadly for text classification are designed. Most likely, the Light Gradient Boosting Machine (LightGBM) (Rufo et al., 2021) and the Extreme Gradient Boosting (XGBoost) (Parente, 2021) have the ability to offer exceptional performance recently.

The Convolutional Neural Network (CNN) (Cheng et al., 2020) model was developed to address the text classification problem for the first time, and it has the highest number of references among deep learning models. The Bidirectional Encoder Representation from Transformers (BERT) with giving consideration to its benefits on various text classification data sets, has been used in designing of text classification models while not designed for text classification tasks (Deepa, 2021).

**Figure 2.** *Different Computer-Based Text Classification Methods*



**Traditional Method.** Data is becoming a tool and fuel for businesses to get crucial insights and boost performance, and data science has dominated nearly all global industries. Data scientists or analysts manually modeled large amounts of data to acquire and process all raw data. Manual text classification includes human annotator that one or more people interpret the content of the text (e.g. an expert in that field) and classify the text based on that. This approach can produce effective results, but it is a time-consuming, expensive, and laborious operation.

Prior to the 1990s, manual or manual procedures accounted for the majority of the common work done for classification (in the sense that an expert is used to decide on the classification of texts). This method of classification has problems such as extracting complete and suitable rules for text classification. These guidelines must be manually created by a knowledge engineer and an expert. This is despite the fact that if it needs to be changed, two persons must update the set again in collaboration, new set is required in which case a different expert is required.

Traditional models broaden the scope of traditional usage while speeding up text classification with better attention. The initial step is to preprocess the raw input text, which typically entails data cleaning, word segmentation, and statistics for traditional model training. By employing characteristics including syntactic depth, sentence length, difficult words, vowels, nominal quotient, qualities, and phrase length, experts evaluated the score using the usual method. The manual labeling of texts used in the old approach is time-consuming, costly, and necessitates the expertise of specialists (Yuan et al., 2020). Due to manual labeling by people, the application of taste and experience causes errors. If several people want to label, the possibility of inconsistency and difference in labeling will arise. The machine learning method is one of many serious methods developed to overcome the problems with the traditional method (Eom & Zhang, 2004).

The readability of a written document can be assessed using readability formulas, which are often employed. A mathematical equation taken from regression analysis is a readability formula. Finding an equation that best captures the link between two variables—the difficulty encountered when reading a book and its linguistic characteristics—is therefore important in order to produce readability formulas (McLaughlin, 1969). In this regard, readability formulas are a way to categorize texts according to how easy or difficult they are to read based on the text's structural elements (Çetinkaya, 2010). Word lengths, word repetitions, vocabulary, the quantity of challenging words, average sentence lengths, the clarity of the text's primary idea, the usage of metaphors, and the difficulty of the grammatical structure must all be taken into account in order to create a valid readability formula (Al-Tamimi et al., 2014). methods for determining readability; they are based on word length, sentence length, the complexity of word use, and There are numerous readability formulas, each with its own unique characteristics and applications (Freda, 2005).

These include the Ateşman Readability, Dale-Chall Readability, Flesch Reading Ease, Flesch-Kincaid Education Level, SMOG Readability, Gunning-Fog Indicator, Fry Readability Graph, Ateşman Readability and Çetinkaya-Long Readability formula. The English Flesch-Kincaid and Gunning-Fog indicators, the Swedish and Danish LIX index, the Spanish Fernandez-Huerta index, and the French Kandal and Moles formulas are often employed (Al-Tamimi et al., 2014). Commonly used formulas for Turkish were developed by Ateşman (1997) and Çetinkaya (2010).

In traditional readability formulas, level determination is made based on only text-based features, and only the quantitative variables such as (word number and sentence length) are built while qualitative variables namely, the intelligibility and comprehensibility part were not almost noticed. While level assignments produced using program data-based formulae take reader characteristics and additional text-based variables into account. According to Collins-Thompson (2014), traditional readability formulas consider the superficial features of a text, they do not take into account the deep structure features such as coherence, cohesion, syntax ambiguities, figurative uses and idioms. In other words, only text features are highlighted.

A classic readability algorithm calculates a grade-level score by calculating the average word and sentence length. According to the majority of experts (Richardson et al., 2012). "not a very specific challenge as the grade level can be very vague" (Richardson et al., 2012).

The readability of a document has been calculated using a variety of formulas. Dale-Chall readability formula (Dale & Chall 1948), Flesch readability formula (Flesch 1948), FOG index readability formula (Gunning 1964), Fry readability graph (Fry, 1965), Senter and Smith,

ARI (1967), McLaughlin, Smog (1969), and Ateşman readability formula are among the popular readability formulas (1997).

1. Dale-Chall (1948): The Dale-Chall readability formula was developed by Edgar Dale and Jeanne S. Chall to evaluate health education materials. This formula calculates readability both by using sentence structure and the ratios of unknown words in a written text containing 3000 frequently used words (Badarudeen and Sabharwal, 2010). In other words, this formula is based on sentence length and unknown (difficult) word count.

The Dale-Chall readability formula steps are as follows:

- About 100 words are taken from every 10 pages of a book. 100 words are selected from every 2000 words for an article. If it is a written text of 200-300 words, the entire text is examined.
- The number of words in the selected sample is determined.
- The number of sentences in the selected sample is determined.
- The number of unknown (difficult) words in the selected sample is determined. Unknown words are words not included in the list prepared by Dale.

2. The Flesch (1948) Ease of Reading: (FRE) Readability Formula was developed by Flesch. The formula measures how complex a text is. The application steps of the Flesch readability formula are as follows:

- The first 100 words are selected.
- The total number of syllables in 100 words is calculated and the number of syllables is divided by the number of words (100) to find the average word length of the text.

The number of sentences in the text is determined by dividing the total number of words by the number of sentences, and the average sentence length is determined (100). The formula used to determine a word's difficulty uses the number of syllables in each word. Each sentence's words serve as an indicator for the syntax's complexity (Stockmeyer, 2009). The Flesch Reading Ease Score, which runs from 0 to 100, is the outcome of the formula. As a result, 0 denotes that the written text is virtually impossible to read, or unreadable; 100 indicates that the readability of the written language is extremely simple (Badarudeen and Sabharwal, 2010).

A basic English text was given a readability score of 60 by Flesch. The majority of documents should have a FRE score between 60 and 70 according to the Microsoft office program (Stockmeyer, 2009), and legal correspondence typically uses it (Savla and Martino, 2012).

3. Gunning, Fog (1952): The Gunning-Fog indicator was developed by Robert Gunning (1952). This indicator focuses on word lengths. While considering the total number of syllables in a text in the Flesh reading ease formula; In the Gunning-Fog indicator, 3 or more syllables in a word are taken into account (Ott and Hardie, 1997). In this sense, the Gunning-Fog indicator includes word length and sentence length as variables (Bezirci & Yılmaz, 2010). It indicates which age group a text is suitable for and in this way gives information about the ease or difficulty of the text (Bezirci and Yılmaz, 2010). According to this indicator, polysyllabic words make a text difficult to read (Ott and Hardie, 1997). A readability score above 12 is defined as a disabled level (Çetinkaya, 2010).

The formula is used to measure the readability of best-selling newspapers and magazines around the world. The first 100-word part is determined and the total number of words and sentences in this part is calculated. Short sentences have better readability level according to the formula. According to the Fog test, it is a difficult text over 12+ points. The most appropriate readability score is indicated as 7 and 8 (Dubay ,2004 and Çetinkaya, 2010).

4. Fry readability graph (Fry, 1965): Fry readability graph is based on two variables, sentence length and syllables (Gilliam et al., 1980). This readability tool is easy to implement and can be implemented without the need for computers or program and is also suitable for very long texts (Badarudeen and Sabharwal, 2010). The Fry readability graph is used as follows:

- Three 100-word parts are selected from the beginning, middle and end of the text under consideration. All proper names are skipped.
- The sentences in each of the 3 selected sections are counted and the average of the sentence numbers is taken.
- The number of syllables in each of the 3 selected sections is found and the average of the syllable numbers is taken.
- The average number of sentences and the average number of syllables in each selected section are marked on the graph. Most of the points fall right next to the curved line. The vertical lines in the graph separate the regions of the readability levels (Fry, 1968).

5. Senter and Smith, ARI (1967): The ARI formula consists of the following stages:

In the first stage, the total number of letters in the text is divided by the number of words and multiplied by 4.71, in the second step, the total number of words in the text is divided by the total number of sentences and multiplied by 0.5 and at the end the results obtained in the first

two stages are added and 21,43 is subtracted from this result. The resulting refers to the readability of the text.

6. McLaughlin, Smog (1969): According to the smog formula, first, 10 sentences are taken from the beginning, middle and end of the text. Sequences of words ending with periods, question marks, and exclamation points are considered sentences. Words consisting of +3 syllables in this section of 30 sentences are counted. If words with +3 syllables are repeated, they should also be counted. The square root of the total number of words with +3 syllables is estimated.

7. Ateşman (1997), Readability Formula: Ateşman arguing that the readability formulas developed for other languages cannot be used for Turkish, adapted the readability formula developed by Flesch into Turkish. This developed formula is the first study in the literature to classify the readability levels of Turkish texts (Çetinkaya, 2010). For the new formula, Ateşman first examined the written texts in different fields and calculated the coefficients in the Flesch reading ease formula for Turkish. In addition, values related to average word and sentence lengths in Turkish were revealed in the study. The formula, which was developed by Ateşman (1997) and used the average word length as syllables and the average sentence length as words, is given in Equation 10. The readability number obtained by Equation 10, which indicates the readability of a text, ranges from 1 to 100 and correspond to “very difficult” and “very easy” readability levels, respectively.

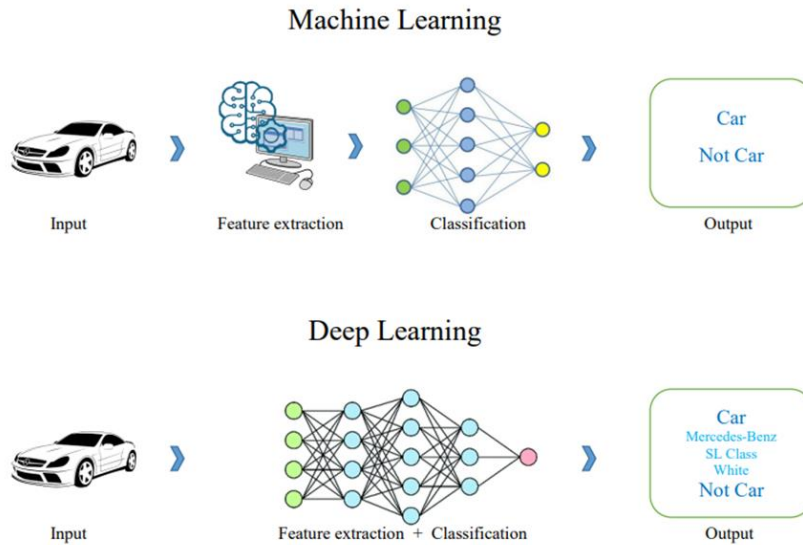
In all of these traditional readability formulas, level determination is made based on only text-based features, while in machine learning formulas, a level assignment is made that includes reader characteristics as well as more text-based variables. The existing readability studies were all developed using the traditional method, and it was found that almost all of the formulas were based on research findings that factors like semantic, syntactic, and lexical difficulty affect the difficulty level of a text but ignored the qualitative variables, that is, the intelligibility part, on which the formulas were built (Graesser et al., 2011; Rubenstein & Aborn 1958). Because they are based on the superficial structure of the texts based on word and sentence length, formulas that measure readability of a text act on texts without visual difficulties. It is not possible to quantify the readability of the text mathematically in terms of semantics. Due to this, readability is typically employed to determine a text's reading complexity using specific mathematical formulas (Ülper, 2010).

These Formulas tend to combine easy units of measure such as sentence length, word length, and word frequency. Texts with shorter sentences, shorter words, and more frequently repeated words; are considered more readable or less difficult texts than texts containing long

sentences, long words, and infrequently used words (Benjamin, 2012). However, as the amount of variables related to the formula increases, the application of the formula will become more difficult (Ateşman, 1997). Although traditional formulas offer many models and grids in such measurements, and the grid diversity is very rich, the high margin of error in the measurements and the fact that the classifications do not reflect the fundamental nature of the text showed that there is a need for more innovative systems with intelligent systems (Hancke et al., 2012; Kanungo & Orr, 2009). They take some time to implement as a methodology, require two or more raters to assure validity-reliability between raters, and occasionally, due to the complexity of the equation utilized, the data is only available to experts.

**Machine Learning.** New ideas for readability text, natural language processing (NLP), and machine learning (ML) were created by Francois and Fairon (2012) as a result of the traditional formulas' focus on surface structure and disregard for complex features like word frequency, lexical meaning, text morphology, and parse tree depth. Machine learning approaches have proliferated not only in the field of text classification but also elsewhere as the capacity and resources for information processing have risen. Machine learning techniques are now used to classify nearly all of the systems in use today. One of the key disciplines of the twenty-first century is machine learning. Pre-artificial intelligence, also known as conventional and post-artificial intelligence, is the focus of these works in the topic of readability. Studies on traditional readability, according to Chen and Meurers (2019), concentrate on sentence length; new or modern studies, however, concentrate on NLP, or natural language processing and machine learning, which emerged with the idea of machines that can think like humans and have gradually begun to be used in all areas of life over time. The system's most crucial aspect is that it is based on a logic that can identify problems, update itself according to changes, and produce results that are independent of people or time.

**Figure 3.** *The Difference Between Machine Learning and Deep Learning*



Text categorization and clustering, document summarization, customer relationship management, web mining, sentiment analysis, and other information retrieval issues have all been successfully addressed using machine learning. Machine learning is one of the most important studies of the 21st century. These studies in the field of readability are divided into pre-artificial intelligence, that is, traditional and post-artificial intelligence.

Understanding readability formulas, the variables they contain, and how to recognize all the extraordinary variants of structural elements in comprehension texts are crucial components of a machine learning-based readability formula. It is feasible to assess a text's readability and intelligibility simultaneously using a readability formula based on data-driven methodologies (Machine learning-based). Although just the text's surface structures are used in the formulas for text readability, more recent readability approaches have begun to incorporate reader factors as well.

Finding all variations of each variable is the first stage in creating a system. The expectations of language in speech have led to an unparalleled rise in computer systems, according to Huawei engineer Qun (2019), and natural language processing (NLP) research is still in its infancy. Whatever is defined as a variable in the computer system, the computer will choose accordingly. An artificial intelligence-based readability formula must carefully consider the definition of the readability formulas, the variables that make up these formulas, and the determination of all the exceptional versions of the text's attributes. Fry (2002) asserts that while the readability formulas provide information about how the text will be scored objectively, other factors such as whether the text's content is appropriate for the target age group, whether the text's illustrations accurately depict the story, how many words are on a page, whether the readability levels are correlated with teaching strategies, and the number of

repetitive sentences and words should also be considered. It doesn't provide an answer to issues like what the rate is. According to Zorbaz (2007), the readability formulas only provide information on the text's structural reading difficulty level; however, when the readability is combined with the qualitative elements of the text, a clear conclusion can be drawn (such as the suitability of the text to the reader, intelligibility). "The programming of computers utilizing sample data and prior expertise to optimize a performance metric" is the definition of machine learning (El Naqa & Murphy, 2015). A classifier is created using machine learning, and it trains on a set of documents whose class labels are manually assigned by a subject matter expert. The classifier then automatically learns the characteristics of which documents should belong to which class.

It is possible to classify texts manually or automatically. Machine learning-based text categorization learns to categorize text based on prior observations rather than depending on manually established criteria. By employing pre-labeled examples as training data, machine learning algorithms can learn different connections between text and select a suitable label for a specific input.

"Tag" refers to the intended class or category into which any input text may be categorized. The automatic method also refers to the use of machine learning and natural language processing to automatically categorize the desired text. This approach is quicker, more cost-effective, and more precise. Feature extraction is the initial step in the machine learning training process for a text classifier. Because the computer is unable to comprehend text the way we do, a method is employed in this process to convert each text into a numerical representation in the form of a vector.

The Bag of Words method, which displays the frequency of a word in a predetermined dictionary, is one of the popular techniques. Contrary to rule-based systems, text categorization based on machine learning is typically significantly more accurate. Additionally, machine learning makes it easier to maintain the classification system and allows us to continuously add new labels to the samples and train the model with them. Machine learning and information retrieval also fall under the category of science that classifies texts. These restrictions occasionally apply to text mining and knowledge extraction from texts as well. In this method, an expert categorizes a set of texts manually, and then by observing the characteristics of these categories, the categorization is done. The process of inference by gathering information from texts is able to categorize a new text that has not been seen before, it is generally a supervised learning method. This method is superior to the traditional method in that it focuses on the

construction of primary categories rather than classification, with the categories being implemented automatically. Sources are already categorized in this method.

Today's categories are extremely careful and based on machine learning methods. Text management based on this principle includes categorizing texts. This method shares some restrictions with information retrieval, and it is occasionally employed in categorization. The study of algorithms and statistical models employed by computer systems that carry out tasks using patterns and inferences rather than explicit instructions is known as machine learning, or ML (Ardabili et al., 2020). It is the science of teaching computers a specific topic without the need for an explicit program. Machine learning algorithms, a subset of artificial intelligence, create a mathematical model based on test data or training data to make predictions or decisions without overt programming. Their main objective is to enable computers to automatically learn and advance without human intervention and explicit programming to adjust their actions appropriately. The goal of machine learning is to create program that can acquire data, learn from it, and apply it in a variety of domains, such as linguistics, engineering, and business. To uncover a pattern in the data and base judgments on it, the learning process starts with observations or data, such as examples, direct experiences, or recipes. There are now numerous facets of daily life where machine learning is used. For instance, machine learning is used by search engines like Google and Bing because their machine learning program knows how to rank a web page. Similarly, Facebook or Apple's photo program that recognizes people's images is a type of machine learning. Spam filtering in e-mail is also one of the applications of machine learning.

Every day, a wide variety of machine learning algorithms are created, and the majority of them are classified according to their learning styles (i.e. supervised learning, unsupervised learning, semi supervised learning). They are categorized based on form and function (such as classification, regression, decision tree, grouping, deep learning, etc.). Regardless of learning style or performance, all machine learning algorithms are as follows:

- I. Representation: A set of classifiers or a language that the computer understands.
- II. Evaluation: Also known as objective performance/scoring.
- III. Optimization: search method; Often the classifier with the highest score.

The fundamental goal of machine learning algorithms is to generalize learnings beyond the trained examples, that is, to successfully interpret the data. Machine learning algorithms are mainly categorized into two types, supervised and unsupervised. A supervised machine learning system can forecast the future using both previously learned lessons and newly tagged data. The learning algorithm develops an inference function to make predictions about the

output values starting with the analysis of the training data set. This type of system is able to identify a target for any new data after sufficient training. It can also compare its output with the correct and predetermined output and modify the model accordingly if any errors are found. When a new document comes to this classifier, the classifier determines the class of the document according to the features it has learned. In terms of machine learning, the classification problem is the "supervised classification" method.

In contrast, unsupervised machine learning algorithms are used when the information required for training is neither classified nor labeled. This type of system does not specify the appropriate output and can only explore the data and infer hidden structures from the unlabeled data.

The semi-supervised machine learning algorithm lies between the previous two types. This system uses both labeled and unlabeled data types for training. Systems that use this method can significantly improve learning accuracy. We usually choose this type of learning when the labeled data obtained need skilled and relevant resources for teaching and learning. Otherwise, accessing labeled data usually does not require additional resources.

Algorithms for reinforcement machine learning interact with their environment through their activities to find errors and rewards. The key elements of reinforcement learning are experimentation, mistake detection, and delayed rewards. This kind of learning enables program agents and machines to autonomously identify the best course of action to enhance performance. A reinforcement signal is used by this system to determine which behavior is preferable through straightforward reward feedback. Massive amount of data can be analyzed thanks to machine learning. Although we might need more time and resources to train it, this learning usually provides quicker and more accurate results when detecting profitable or risky options. A combination of machine learning and artificial intelligence and cognitive technologies can be effective in processing large amounts of information.

Machine learning is designed to help computers and systems improve their efficiency by increasing data as they perform the desired task. By receiving rewards and punishments, biped robots are taught to walk by automatically recognizing several examples of the desired face. Machine learning research spans a wide spectrum of topics. While some researchers are interested in applying machine learning techniques to novel issues, others are interested in developing new learning methods and researching the viability and quality of such approaches.

Machine learning significantly increases the pace of data analysis while reducing operating costs. For instance, machine learning is used in the oil and petrochemical industry to measure operational data from all drillings and analyze the data to modify algorithms for better

and more optimum extraction in subsequent drillings. Machine learning has also usage in human's daily life such as social networks, online shops, in search engines and advertising, mobile locks, personal assistants, speech recognition and traffic forecasting. Making predictions is a common element of all the machine learning applications that are being developed today, and their number is growing daily. These predictions can be made in various fields in which artificial intelligence is capable. Image recognition, sound recognition, trend recognition, and the relationship of different elements within a data set are among these. These apps' main goal is to enable computers to finally carry out human duties more accurately and securely.

### ***Text Classification Studies on English Texts***

In many Natural Language Processing (NLP) applications, such as sentiment analysis topic labeling, question answering, and dialog act classification, text classification is an essential task. Trying to process and classify large amounts of text by hand in the era of information explosion is both time-consuming and challenging. Human factors, such as tiredness and expertise, can also affect manual text classification accuracy. In order to achieve more reliable and less subjective results, machine learning methods should be used to automate the text classification procedure. Furthermore, this also improves the efficiency of information retrieval and alleviates information overload. Classification of scientific texts can be effective in creating expert systems and better search for articles and researchers have easier access to the articles available in all scientific databases. ranking texts are important in many ways. Rapid increase of internet websites, users, presentation Online storage providers and social media, the amount of available data, especially the publication of data Increases science daily. Feature extraction is an important step in text classification and the most important application of text classification is the ability to separate scientific articles into different classifications. It can be done in the classification of conference articles, theses classification, keywords Automatic considered for the bank of scientific articles. Most text mining methods use search based on terms. All these methods are affected by common challenges such as multiple words with one meaning and one word with multiple meanings. Patterns have many advantages over other term-based methods. Explore the rankings Pattern can be useful in increasing patterns. In addition to solving common problems in searching based on terms is effective. The need for text classification has grown so dramatically that it cannot afford manually. Because the documents of libraries, patent offices, public hospitals and Other institutions are rapidly becoming electronic each year, so it is possible to rank them It cannot be done by hand. Text classification on document cataloging issues for creating archives sorted, automatic

classification of documents for automatic routing systems and data filtering unrelated flow is very important for news agencies. With the growth of electronic documents on the Internet today, it is essential for scientific systems and content systems to automate text classification. Automatic text classification can be interpreted as the task of assigning predefined categories to textual documents. Many information and management organization tasks are performed prior to the operation of a prerequisite process such as text retrieval, sorting, subject identification, filtering and searching, automatic text sorting, several machine learning algorithms are proposed for text classification. The K-Nearest Neighbor algorithm is one of the most successful machine learning algorithms in this field. However, the classification of new cases is one of the limitations of this algorithm which suffers from high computation. Feature selection techniques are one of the methods available to address this challenge. There are various ways of categorizing scientific texts in the literature. Some of the work done in this area can be text mining, document vector correction, natural language processing, linguistic structure identification, categorization of texts, feature vector optimization, jaccard similarity. But this field can have a great deal of research due to the comprehensiveness of the subject and the exceptions such as the style of literary, artistic, interactive and scientific writing and, on the other hand, the language of the texts (Chen et al., 2020; Chen et al., 2019; Kesavan et al., 2017, Lee & Demoncourt, 2016; Liu et al., 2015; Zhu et al., 2015; Tai et al., 2015; Kalchbrenner et al., 2014).

## SECTION THREE

### Methodology

#### Research Method

The study is a quantitative study in which the text classification scorings made by traditional formulas and field experts on the same sample (YDS reading-comprehension texts) were compared with the text classification scoring made using machine learning algorithms from modern approaches, and this scoring comparison will evaluate the use of machine learning algorithms in text classification. In this context, the main reason for choosing a comparison-oriented study is to reveal the differences and similarities between the measurements and to offer an in-depth perspective on the use of machine learning algorithms that include advanced technologies in selecting the appropriate text in the measurement of English reading comprehension skills (Azarian, 2011; Pickvane, 2005). In particular, the basic design of the study is called as variation-finding used to determine the logical differentiation between the measurements for a specific phenomenon (text readability-difficulty across scores of different assessments techniques), which is one of the four designs presented by (Tilly 1984, p. 82). Basically, the study moved on compare-contrast based descriptive research design seeking to compare and contrast two scorings through the description of identified systems (Fraenkel et al., 2012; Gay et al., 2012; McMillan & Schumacher, 2014). The main advantage of this design is to provide systematic information about the text classification phenomenon through machine based and human based measurements (Ethridge, 2004, p.24). Therefore, systematic collection of information requires careful selection of the data source to be studied and careful measurement (Fox & Bayat, 2007, p.45). The other advantage of this research design is that it provided the possibility to observe the both machine learning algorithms and grid based human scoring in a completely natural unchanged environment.

#### Data Source

##### *Population and Sample*

The universe of the participants who will participate in the scoring part of the classification measurement grid-oriented classification, which is based on traditional formulas, consists of expert academicians in the English language teaching department in Turkey. Based on the purposeful random sampling technique (Fraenkel et al., 2012; Gay et al., 2012; McMillan & Schumacher, 2014), the study sample will consist of academicians who are assessing and

evaluating specialists in English language teaching departments. Because the classification of the reading-comprehension texts by measuring the readability or difficulty in assessment is the focus of the study, filling the classification measurement grid based on traditional formulas by English assessment experts will reduce the doubts about why specific situations are selected and will ensure the reliability of the research (Fraenkel et al., 2012; Gay et al., 2012; McMillan & Schumacher, 2014).

### ***YDS Corpus***

The text-oriented universe of the study for machine learning algorithms consists of the reading-comprehension texts in the foreign language knowledge level examination (KPDS-ÜDS-YDS) conducted by the Measurement, Selection and Placement Center (ÖSYM) to determine the level of English language knowledge. The study sample is the reading-comprehension texts in the Foreign Language Proficiency Exams (YDS) held between 2013-2022, based on corpus-oriented criteria. In this context, the main reason for taking the exams between 2013-2022 is that the language placement exams made in 2013 with the decision taken by the Assessment, Selection, and Placement Center were gathered under a single roof as the Foreign Language Proficiency Exam (YDS). The first of the main criteria for focusing specifically on YDS exams is; Since high-stakes testing, that is, exams with important results for the participant, have important effects for the public and policymakers, the readability/difficulty levels of the texts in these exams must be focused on valid and reliable assessment-evaluation techniques and measurements (Stobart & Eggen, 2014). Another important criterion is that in corpus-oriented studies, the use of archived corpora in databases is recommended due to the possibilities it provides in data entry in program-oriented analyzes (Maqsood et al., 2020). The third criterion is being “suitable for the scope of comparison”, which includes six basic sub-criteria (Moreno, 2008). In this context, since the nature of the study is based on comparison, form, channel, dialectal value, tone, format, and language code can be used to classify text readability through both machine learning algorithms and traditional formulas. (language code) is also taken into account and to be explained in order for this study;

- The texts are in written form
- Electronic versions are essential as access channels
- The dialectical value is that it is the official language
- It is formally oriented in style
- In format it is in exam format

- The language code is also English. Preferring English is the selection of reading-comprehension texts suitable for the level of English education, as well as the validity-reliability of the tokens for English text classification in both traditional and machine learning algorithms (Tseng et al., 2019).

The text-oriented universe of the study, which will be classified with traditional formulas, consists of the reading-comprehension texts in the foreign language knowledge level examination (KPDS-ÜDS-YDS) conducted by the Measurement, Selection and Placement Center (ÖSYM) to determine the level of English language knowledge. The sample selected for this study step will be determined by using different sampling techniques. First of all, as in machine algorithms, the reading-comprehension texts in the Foreign Language Proficiency Exams (YDS) held between 2013-2022 will be the texts to be classified in traditional formulas, and in this context, the reasons expressed for machine learning algorithms are based on traditional formulas and is valid in the classification step (Maqsood et al., 2020; Stobart & Eggen, 2014). At this point, in order to determine the texts to be classified by assessment and evaluation experts, it will be expected that the machine learning algorithms classification process of the reading-comprehension texts between the years 2013-2022 will be completed; Based on the classifications that will emerge as a result of this measurement, the texts will be divided into groups according to years and difficulty levels. As the next step, the stratified sampling technique will be followed for selecting texts to cover both the year and difficulty levels and thus represent the texts between 2013-2022. Because with this technique excludes the selection for the year and difficulty levels from the risk of random distribution, and an equal selection will be realized (Dawson & Trapp, 2001). In addition, researcher bias will be controlled since the plates will be chosen by drawing lots (McMillan & Schumacher, 2014). Finally, selection processes will be carried out by determining the number of texts to be used for traditional formulas by following the sample size formula suitable for the nature of the study to represent and compare the reading-comprehension texts between the years 2013-2022.

## **Data Collection Techniques/Tools**

### ***Grid***

One of the measurement tools that will be used in data collection in the study is the measurement used in the classification based on traditional formulas, which was developed as a result of third-generation studies in traditional formulas (Appendix 1). In this context, different grids are available in the literature according to various traditional formulas. The main reason for various measurement grids is that, as a result of the studies, the grids were insufficient

to meet the relevant parameters in text classification and the development of different grids by structuring new formulas (Maqsood et al. 2020). First-generation formulas in this area examine text readability/difficulty based on structural features, in other words average sentence length, number of syllables, and total word count (Rush, 1985; Danielson, 1987). However, since the structures discussed in these formulas remained superficial, they excluded topological elements such as cohesion and coherence that determine legibility in the texts, and therefore the text was insufficient in classification (Roman et al., 2021). Second-generation formulas were developed based on this shortcoming and included higher-level text dimensions such as macrostructures and text density. However, they were also insufficient in classification as they were factors affecting the readability level of meta-cognitive patterns as well as cognitive structures in texts (Maqsood et al., 2020). Based on the deficiencies in the first and second-generation formulas in the literature and based on classification with parameters similar to machine learning parameters in order to be able to compare with machine learning algorithms, the most appropriate traditional formula-based grid is the grid that emerged in the third generation studies in traditional formulas (François & Fairon, 2012). Because in the third-generation traditional formulas, natural language processing is patterned over artificial intelligence, and both cognitive and meta-cognitive parameters are activated. In other words, it includes a holistic analysis of syntactic, semantic, lexical, and topological features with a view of the interaction between parts.

### ***Machine Learning Algorithms***

The text classification process of the machine learning algorithms system is defined as training in the literature, where the operational steps are entered into the system through parameters called “tags.” Thus the different relationships between text pieces and the expectation of a certain output (ie tags) for a certain input (i.e. text) are commanded through algorithms. It consists of the training step, called prediction, and the estimation step, which is expressed as a prediction in the literature, where classifications emerge as a result of the processing and analysis of the relevant parameters through algorithms (Petersen & Ostendorf, 2009). The basic working principle of this system is that it has the feature of converting each text into a digital representation in the form of a vector and making various analyzes over these vectors. In this context, the processes followed during the data collection phase are as follows (Ortego & Martínez Sánchez, 2019):

1. Text content collection (Pre-processing process) = Determination of the collection and entry into the system.

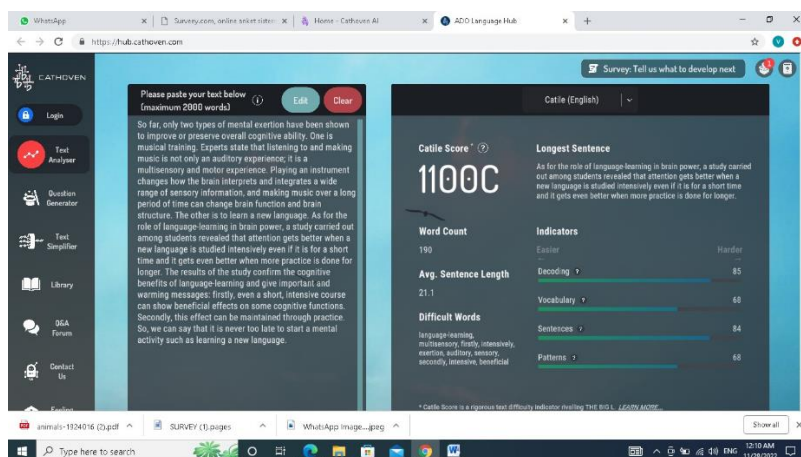
2. Coding the parameters (Training) = Coding the English parameters defined in the literature into the system. Coding the syntactic, semantic, topological, and lexical parameters into the system
3. Vectorization of texts, testing of vectors, and classification (Prediction)
4. Entering the scoring made by the assessment and evaluation experts into the system and comparing the vectors of machine learning algorithms and expert vectors.

### **Data Collection Process**

The study process was managed in three stages before the data collection process, data collection process-analysis, and reporting. Before the data collection process, first an ethics committee report was obtained from the Atatürk University Institute of Educational Sciences ethics committee for the study, and then the necessary permissions were obtained from OSYM for the use of YDS texts in the study. After the necessary permissions were completed, the data collection process began at the second stage of the process.

The first step of the data collection-analysis process was planned to form the corpus of YDS reading-comprehension texts, which is stated as pre-processing in the literature and is defined as the pre-processing step, in the focus of the aim of the study and the universe-sample processes and criteria. After the related corpus was compiled, the related tokens were looked for in the program by defining the syntactic, semantic, topological and lexical parameters defined for detecting English text readability levels in the literature for machine learning algorithms (Kumiko et al., 2010). Then, CATHOVEN program was chosen and the corpus was uploaded to the system and the relevant entry was completed. On the basis of the established training in the program, in which the expectation of a certain output (that is, labels) is commanded through algorithms, CATHOVEN started prediction, which is expressed as prediction in the literature, where the classifications emerge as a result of processing and analysis of the relevant parameters through algorithms, and the prediction step, which is related to machine learning algorithms. The first step was completed (Petersen & Ostendorf, 2009).

**Figure 4.** *The Cathoven Software View*



The second step of the data collection-analysis process was to send the texts determined by the appropriate sampling methods among the classified texts to the assessment and evaluation experts for classification by scoring with the classification grid selected from the literature. After determining the e-mail addresses actively used by academicians who are experts in assessment and evaluation in the field of English teaching, one of the systems with a user-friendly interface from the online survey systems, in which the ethics committee and other permissions of the study are added, was identified as a result of the examination with the consultant and the link of the grid organized on this system will be shared. Invitation mail will be sent to them. In the relevant link, the Participant Consent Form and Information Text (Annex-2) will be located on the first page, ensuring that the participants' privacy and rights are protected and that the participation is voluntary and provides detailed information about the study. The second page of the link will contain a short informative text on which parameters and the scope of the scoring process on the grid, thus eliminating the uncertainty of how the participants will score.

The last step of the data collection-analysis process was to enter the scores from the CATHOVEN assessment and evaluation experts, the scores of which were converted into the scores of CATHOVEN, into SPSS. Then, machine learning algorithms and traditional method classifications was compared, the data obtained was reported in the results and discussion section.

## **Data Analysis**

In the data analysis process, CATHOVEN program will be used to compare the classifications of texts with modern data-based methods and traditional formulas with modern data-based machine learning algorithms classifications. In this context, the algorithms to be followed for the realization and comparison of the estimation step, which is expressed as a

prediction in the literature, where the classifications arise; in other words, the analyzes are as follows: Penalized logistic regression (glmnet), linear discriminate analysis (lda), Naïve-Bayes (nb), partial least squares regression (pls), support vector machine radial basic function kernel (svmR) and random forest (rf). The main reason why CATHOVEN program is preferred in the analysis process is that it provides the opportunity to provide cross-validity by comparing all algorithms in classification by using various algorithms, that is, analysis models (Petersen & Ostendorf, 2009). According to the working principle of the program, after the first five of algorithms perform the classification, all classifications are compared and tested with the last algorithm (Chang & Lin 2011). One of the main reasons why the related program and machine learning algorithms are reliable in classification is that each analysis is repeated five times within itself (Sung et al., 2015).

As for comparing and contrasting experts scores and CATHOVEN, paired samples T-test was used because the scores of experts were converted into the scores of CATHOVEN and so two paired sample scores were established for analysis. Besides, Cronbach alpha scores were measured for both CATHOVEN and experts' scores and shared in results and discussion part.

### **Researcher Role**

In the study, the main starting point for the researcher to adopt comparison-oriented quantitative is the rationale for ensuring objectivity and minimizing the bias in the data. In this context, in the data collection process with the grid, which will be sent to the participants in the English text classification developed over the traditional formulas, which is one of the stages of the data collection process, the researcher will follow the process by mail regardless of the location, thereby eliminating the researcher's effect. Thus the researcher bias will be controlled. Also, in classifying texts with machine learning algorithms, the researcher's impartiality will be ensured, and the situations where the researcher's subjectivity will be discussed will be controlled.

### **Validity and Reliability**

For the internal validity of the study, first of all, the literature on traditional methods of text classification in English and machine learning algorithms was examined in detail, and a conceptual and theoretical framework was created on the subject. Based on the conceptual and theoretical framework created, the rationale for which texts would be studied was structured based on the literature, and machine learning algorithms were determined with the grid based on traditional formulas that can be used in classification. The thesis advisor's opinion was taken after confirmation by a program engineer academic who is an expert in this field. The grid to

be used by assessment and evaluation experts in the study is a grid developed to classify the readability level of English texts in the literature (Maqsood et al. 2020). It is a measurement tool that was structured as a result of the studies of English assessment and evaluation experts and English language experts who are experts in the related grid field, and measurements confirmed its validity and reliability in the universal context and its usability in different contexts has also been tested. In this study, the validity and reliability of the relevant grid was tested with a preliminary pilot study and its contextual validity was confirmed. In the data collection process to be carried out over the grid, before the online grid was shared with the measurement and evaluation experts; the details of the study (reason and importance), the features of the scoring grid, and conceptual information was provided, thus ensuring voluntary participation and trust in the study (Fraenkel et al., 2012; Gay et al., 2012; McMillan & Schumacher, 2014). This detailed information process ensured that the data to be obtained are real data. In order to establish the study's external validity, the research process and the applications to be made in this process -such as the research design, data collection tools, analysis and interpretation of the data- were presented in the study in detail. For the external reliability of the study, all processes, including data collection and analysis, were explained in detail and kept by the person who made the study so that it can be accessed by someone else in case of need (Fraenkel et al., 2012; Gay et al., 2012; McMillan & Schumacher, 2014).

For the internal reliability of the study, it was essential to pay attention to the diversity of the data to be obtained as much as possible and to select the document group suitable for the nature of the comparison in this context (Fraenkel et al., 2012; Gay et al., 2012; McMillan & Schumacher, 2014). Similarities and differences in the obtained data was be reported transparently in all details. In addition, the selection of assessment experts for classification with traditional formulas is among the important dimensions in ensuring the study's internal validity and internal reliability. Finally, since all text readability parameters have been coded into the artificial intelligence program, when it evaluates English texts, it has paid attention to the most detailed issues (which are beyond the abilities of human participants) as a result of the details related to all classification methods having been observed and thus increasing its reliability.

## SECTION FOUR

### Results

In this chapter, quantitative research findings will be presented in relation to English texts' difficulty classification through machine learning algorithms. The findings of the three research questions mentioned previously are presented in this chapter as quantitative findings. As a preprocessing step, nine YDS reading comprehension texts were prepared. Then the machine learning algorithm "CATHOVEN" was selected for prediction. A second test was administered through an online survey sent to ELT experts via email; randomly selected texts were sent to assessment and evaluation experts who were pre-trained on the selecting methods. CATHOVEN assessment experts' scores were then entered into IBMSPSS 23. At the end machine learning algorithms were compared with traditional classification methods. The results of the analysis of findings are demonstrated via tables and quotations.

#### Findings Related to First Research Question

**Research Question 1:** Is there a significant difference between the difficulty classification of YDS reading-comprehension texts with a program using the modern data-based methods (machine learning algorithms) and the classification made over traditional formulas by field experts?

As part of this analysis, field experts used a grid that had been developed with traditional formulas to classify YDS reading-comprehension texts while machine learning algorithms, which are one of the modern approaches, were used to classify YDS reading-comprehension texts. In the study, data mining was used to classify texts using tokens based on syntactic, lexical, semantic, and topological parameters, and the information provided by data mining was used to determine readability and difficulty level (Ortego & Sanchez, 2019). There were two basic methods; one was a grid-based system of scoring teachers on texts using traditional formulas, and the other was intelligent program-assisted classification (cathoven) using machine learning algorithms, one of the most modern approaches to data processing. There are number of readability measures and there is no standard for choosing the readability formulae (Breese & Burman, 2005). For this study, in order to calculate readability scores for each text, eight algorithms were used: Flesch-Kincaid Grade Level (F-KGL), Gunning Fog Index, SMOG (Simple Measure of Gobbledygook) Grade, Automated Readability, Coleman-Liau Index,

Linsear Write, Dale-Chall-Readability, and McAlpine EFLA. On the basis of this algorithms, the academicians' scores were transformed to the scores that would make the comparison of scores by machine learning and academicians of the readability/difficulty classification of texts.

In order to compare the readability level scorings of CATHOVEN and Experts of YDS reading-comprehension texts, the paired sample t-test is used to compare two groups. *p* scores show whethere there is significant differences between the results of modern data based software and the traditional formulas based grid. In Table 2, the t-test results of the difficulty classification scores of CATHOVEN and Experts for YDS reading-comprehension texts are shown.

In general, among eight pairs, the findings show that for five pairs there is a statistically significant difference (Pair 1, Pair 2, Pair 5, Pair 6, and Pair 8) while for the the left pairs (Pair 3, Pair 4, and Pair 7 ) there is no significant difference.

**Table 2.** *T-test Results of the Difficulty Classification Scores of CATHOVEN and Experts for YDS Reading-Comprehension Texts*

|        |                        | n | x     | s.d.  | t     | p     | df |
|--------|------------------------|---|-------|-------|-------|-------|----|
| Pair 1 | Flesch Kincaid Grade   | 9 | 4.11  | 1.364 | 0.847 | 0.004 | 8  |
|        | Flesch Kincaid Grade2  | 9 | 9.22  | 1.922 |       |       |    |
| Pair 2 | Gunning Fog Index      | 9 | 6.22  | 1.563 | 0.861 | 0.004 | 8  |
|        | Gunning Fog Index2     | 9 | 11.11 | 2.261 |       |       |    |
| Pair 3 | SMOG Index             | 9 | 10.00 | 3.041 | 0.461 | 0.211 | 8  |
|        | SMOG Index2            | 9 | 11.33 | 1.871 |       |       |    |
| Pair 4 | Automated Readability  | 9 | 11.75 | 3.041 | 0.160 | 0.682 | 8  |
|        | Automated Readability2 | 9 | 12.33 | 1.803 |       |       |    |
| Pair 5 | Coleman-Liau-Index     | 9 | 13.56 | 3.283 | 0.862 | 0.004 | 8  |
|        | Coleman-Liau-Index2    | 9 | 8.44  | 1.590 |       |       |    |
| Pair 6 | Linsear Write          | 9 | 5.67  | 1.871 | 0.869 | 0.002 | 8  |

**Table 2.** (continuation)

|        |                             |   |       |       |        |       |   |
|--------|-----------------------------|---|-------|-------|--------|-------|---|
|        | Linsear Write2              | 9 | 11.22 | 2.587 |        |       |   |
| Pair 7 | Dale-Chall-Readability      | 9 | 10.00 | 3.808 | -0.232 | 0.548 | 8 |
|        | Dale-Chall-Readability2     | 9 | 9.33  | 0.707 |        |       |   |
| Pair 8 | McAlpine EFLAW Readability  | 9 | 17.33 | 3.742 | 0.938  | 0.000 | 8 |
|        | McAlpine EFLAW Readability2 | 9 | 27.11 | 3.655 |        |       |   |

Pair1: As is seen in the test result in Table 2, the difference between classification scores of Cathoven and experts for YDS reading comprehension texts on the basis of Flesch Kincaid Grade was found significant; the t-statistic was 0.847 with  $df=8$  ( $p < .05$ ). This result showed that the text readability classification by machine learning algorithms and traditional formulas based measurement differs in terms of the difficulty. When the table was reviewed more for Flesch Kincaid Grade, the mean score of experts' YDS reading comprehension texts classification were  $\bar{x}=4.11$  while the mean score of Cathoven was graded  $\bar{x}=9.22$ . It can be understood that, on the basis of Flesch Kincaid Grade, the YDS comprehensions texts chosen are difficult for readability to Cathoven, but for experts they are readable.

Pair2: As is seen in the test result in Table 2, the difference between classification scores of Cathoven and experts for YDS reading comprehension texts on the basis of Gunning Fog Index was found significant; the t-statistic was 0.861 with  $df=8$  ( $p < .05$ ). This result showed that the text readability classification by machine learning algorithms and traditional formulas based measurement again differs in terms of the difficulty. When the table was reviewed more for Gunning Fog Index, the mean score of experts' YDS reading comprehension texts classification were  $\bar{x}=6.22$  while the mean score of Cathoven was graded  $\bar{x}=11.11$ . As in Flesch Kincaid Grade, it can be understood that, on the basis of Gunning Fog Index, the YDS comprehensions texts chosen are difficult for readability to Cathoven, but for experts they are readable.

Pair 3: As is seen in the test result in Table 2, the difference between classification scores of Cathoven and experts for YDS reading comprehension texts on the basis of SMOG Index was not significant; the t-statistic was 0.461 with  $df=8$  ( $p > .05$ ). When the table was reviewed more for SMOG Index, the mean score of experts' YDS reading comprehension texts classification were  $\bar{x}=10.00$  while the mean score of Cathoven was graded  $\bar{x}=11.33$ . It is seen

that, on the basis of SMOG Index, the YDS comprehensions texts chosen are difficult for readability both to Cathoven and to experts.

Pair4: As is seen in the test result in Table 2, the difference between classification scores of Cathoven and experts for YDS reading comprehension texts on the basis of Automated Readability was not significant; the t-statistic was 0.160 with  $df=8$  ( $p > .05$ ). When the table was reviewed more for Automated Readability, the mean score of experts' YDS reading comprehension texts classification were  $\bar{x}=11.75$  while the mean score of Cathoven was graded  $\bar{x}=12.33$ . It is seen that, on the basis of Automated Readability, the YDS comprehensions texts chosen are difficult for readability both to Cathoven and to experts.

Pair5: As is seen in the test result in Table 2, the difference between classification scores of Cathoven and experts for YDS reading comprehension texts on the basis of Coleman-Liau-Index was found significant; the t-statistic was 0.862 with  $df=8$  ( $p < .05$ ). This result showed that the text readability classification by machine learning algorithms and traditional formulas based measurement again differs in terms of the difficulty. When the table was reviewed more for Coleman-Liau-Index, the mean score of experts' YDS reading comprehension texts classification were  $\bar{x}=13.56$  while the mean score of Cathoven was graded  $\bar{x}=8.44$ . On the basis of Coleman-Liau-Index, the YDS comprehensions texts chosen are difficult for readability to experts, but for Cathoven they are readable.

Pair6: As is seen in the test result in Table 2, the difference between classification scores of Cathoven and experts for YDS reading comprehension texts on the basis of Linsear Write was found significant; the t-statistic was 0.869 with  $df=8$  ( $p < .05$ ). This result showed that the text readability classification by machine learning algorithms and traditional formulas based measurement again differs in terms of the difficulty. When the table was reviewed more for Linsear Write, the mean score of experts' YDS reading comprehension texts classification were  $\bar{x}=5.67$  while the mean score of Cathoven was graded  $\bar{x}=11.22$ . On the basis of Linsear Write, the YDS comprehensions texts chosen are difficult for readability to Cathoven, but for experts they are readable.

Pair7: As is seen in the test result in Table 2, the difference between classification scores of Cathoven and experts for YDS reading comprehension texts on the basis of Dale-Chall-Readability was not significant; the t-statistic was -0.232 with  $df=8$  ( $p > .05$ ). When the table was reviewed more for Dale-Chall-Readability, the mean score of experts' YDS reading comprehension texts classification were  $\bar{x}=10.00$  while the mean score of Cathoven was graded  $\bar{x}=9.33$ . It is seen that, on the basis of Dale-Chall-Readability, the YDS comprehensions texts chosen are difficult for readability both to Cathoven and to experts.

Pair8: As is seen in the test result in Table 2, the difference between classification scores of Cathoven and experts for YDS reading comprehension texts on the basis of McAlpine EFLAW Readability was found significant; the t-statistic was 0.938 with  $df=8$  ( $p < .05$ ). This result showed that the text readability classification by machine learning algorithms and traditional formulas based measurement again differs in terms of the difficulty. When the table was reviewed more for McAlpine EFLAW Readability, the mean score of experts' YDS reading comprehension texts classification were  $\bar{x}=17.33$  while the mean score of Cathoven was graded  $\bar{x}=27.11$ . On the basis of McAlpine EFLAW Readability, the YDS comprehensions texts chosen are difficult for readability to Cathoven, but for experts they are readable.

With respect to the first research question, it was found that, there is a statistically significant difference between traditional formulas used by field experts and modern data-based methods (Cathoven). Compared to field experts, CATHOVEN program shows that the in terms of readability the texts are difficult. In other words, reading-comprehension texts shows statistically significant differences between traditional and modern readability formulas in terms of the overall scores.

### **Findings Related to Second Research Question**

**Research Question 2:** What is the reliability score of the difficulty classification scoring of YDS reading-comprehension texts in the measurement made with a program using modern data-based methods (machine learning algorithms)?

Modern data based methods take into account both readability and intelligibility when determining a text's readability unlike conventional formulas of text readability, in which only the surface structure of the text was considered, Accordingly, The CATHOVEN programme was used to analyze not only surface structure, but also complex features such as word frequency, lexical meaning, and text morphology. One of the most important features of this programme was that it was based on a logic that detected the problem, could update itself in response to changes, and had results that did not change with the individual or time (Bonner et al., 2023; Marsh, 2022).

In table 3, The Cronbach's Alpha of Cathoven program is shown. According to the Cathoven results the reliability score of Cathoven program was 0.918 ( $P>0.05$ ), which indicates that the text readability classification of Cathoven is highly reliable. Thus, it can be concluded that modern data based methods are based on reliable scoring in the classification of texts and are seemingly useful tools to be used in the readability classification of English texts.

**Table 3.***The Cronbach's Alpha of CATHOVEN Program*

| Cronbach's Alpha | Cronbach's Alpha Based on Standardized Items | N of Items |
|------------------|----------------------------------------------|------------|
| 0.887            | 0.918                                        | 9          |

As mentioned in the literature review, data-based machine learning algorithm is used to determine readability/difficulty levels of texts, and classifications are made by analyzing elements specific to the language. (Miltakaki & Troutt, 2007). Data-driven methods (machine-learning based), considering additional complex features such as word frequency, word dictionary, text morphology, and the depth of the parse tree, Improved the validity and reliability of the description of the sentence complexity. While, according to Crossley, Allen, and McNamara (2011;88), “traditional reliability formulas are simple algorithms that measure text reliability based on sentence length and word length.” Thus, modern data based methods might be used to develop an effective solution for text analytics, problems involving text classification and clustering, document summarization, customer relationship management, web mining, sentiment analysis, and other information retrieval problems in the future using a variety of modern data-driven methods.

### **Findings Related to Third Research Question**

**Research Question 3:** What is the reliability score of the difficulty classification scoring of YDS reading-comprehension texts in the measurements made by English assessment-evaluation experts through traditional formulas?

This study examined the readability of YDS comprehension texts using the traditional readability formulas mentioned earlier. Therefore, in the research on readability, assessment-evaluation experts were asked to analyze the readability of YDS texts on the basis traditional formulas based grid. In the grid, the criteria put in the literature to analyze the texts were provided to the experts, as follows:

- Syntactic depth: Complex sentences are less readable than simple sentences and the proportion of complex sentences in a text is often used as a feature in traditional readability formulas [Klare, 1963].
- Sentence-length: sentence-length is an indicator of text complexity Melin and Lange [1995].
- Difficult words: the number of words containing more than six letters is calculated as the average number of difficult words per sentence.

- Vowels: the number of syllables per word is often used in readability formulas as a measurement of morphological complexity [Klare, 1963], where more syllables make the words more complex.
- Nominal quotient (NQ): the amount of information in a text influences the readability of the text; a less informative text is more readable than a highly informative one.

In this regard, it is essential to point out that the error, which would from text selection, was reduced by randomly selecting English texts. Moreover, as the respondents of the survey were evaluation-assessment experts in the field of English language teaching, the distribution of the scorings by the experts were expected to be close to each other.

In table 4, the Cronbach's Alpha of English assessment-evaluation made by experts through traditional formulas are given. Through Cronbach's alpha, the survey's reliability was calculated using the participants' data and was found to be 0.469. Since the rate of 0.469 in the test is less than 50 percent ( $P < 0.05$ ), the answers given by the academicians to the questions are not reliable.

**Table 4.** *The Cronbach's Alpha of English Assessment-Evaluation Made by Experts Through Traditional Formulas*

| Cronbach's Alpha | Cronbach's Alpha Based on Standardized Items | N of Items |
|------------------|----------------------------------------------|------------|
| 0.265            | 0.469                                        | 9          |

It is seen that in traditional readability models, comprehension factors such as coherence (Gernsbacher 1997), meaning construction, and cognitive processes such as decoding and syntactic parsing (Perfetti, Landi & Oakhill 2005) are not taken into consideration while cognition factors such as word length and sentence length (Crossley, Allen & McNamara 2011) are accounted. Thus, it can be concluded that field experts' answers regarding the analysis of comprehension texts contain more errors. It is evident that this type of test is prone to human error.

Besides, the use of readability tests has been criticized for being too simplistic and neglecting to take into account many other variables that could affect reading comprehension (Bitgood 1996; Harrison & Bakker 1998). The traditional formula for determining text readability doesn't provide reliable indicators of comprehension, nor does it support cognitive processes associated with comprehension (McNamara & Magliano 2009). The findings of the study imply that while readability formulas are valuable measures for evaluating level of text difficulty, they should be used cautiously. Further research seems necessary to check the validity of the readability formulas and the findings of the present study.

## SECTION FIVE

### Discussion, Implications and Conclusion

A variety of implications will guide the conclusion beyond the findings. The section will conclude with the study's limitations and recommendations for future studies.

#### Discussion

For reading comprehension in foreign language education and language studies, English texts play a central role. This is both in the selection of appropriate assessment and evaluation texts for the level. In addition, they are among the factors that affect course materials selection. In fact, many studies in the related literature show that there is a positive relationship between readability and comprehension of foreign language texts. There are a variety of digital and print resources available for the teaching and assessment of English language in the digital age. There are studies in the international literature that suggest that in this context, smart data-based program following certain formulas, which have become increasingly common in text classification over the last ten years, is replacing expert-oriented traditional classification methods in various disciplines, including education, and classification of English texts. It takes time for traditional formulas to be implemented as a methodology. Because of the complexity of the equation, they often require more than one rater or even more to ensure validity-reliability between raters. Due to this, data-driven smart program was Used to classify English texts as a result of input-output processing in natural language processing using machine learning algorithms. In this case, it is aimed to compare the text classification scoring made by field experts with the grid developed with traditional formulas on YDS reading-comprehension texts and the text classification scoring made using machine learning algorithms. It is a modern approach to analyzing important parameters. This scoring comparison aims to measure and evaluate the effectiveness of machine learning algorithms in text classification based on readability/difficulty levels. Based on a literature review, the study was designed to answer three main research questions with quantitative aspects. The data collection process began by collecting nine YDS reading comprehension texts held between 2013-2022. The ethics committee report for the study was obtained from the Atatürk University Institute of Educational Sciences ethics committee. And OSYM permissions were obtained to use YDS texts in this study. ELT academicians were invited via email to participate in an online survey system to classify and score reading comprehension text. This was after receiving the report

from the ethics committee and approval from the Atatürk University Institute of Educational Sciences, as with traditional text classification.

In the end, machine learning algorithm “CATHOVEN” program was utilized to compare text classification methods with modern data-based methods and traditional formulas with field experts. Using CATHOVEN program in the analysis process was desirable since it allowed cross-validation by comparing all algorithms in classification. Later, the readability level of two variables was calculated as part of the data analysis. Two independent groups are compared using the independent sample t-test. Sig Alpha was calculated using two methods (Cathoven and ELT experts. Readability statistics were calculated for every text and algorithm. Readability scores were calculated for each of the YDS reading comprehension texts using each of the eight readability measures (Flesch-Kincaid Grade Level (F-KGL), Gunning Fog Index, SMOG (Simple Measure of Gobbledygook) Grade, Automated Readability, Coleman-Liau Index, Linsear Write, Dale-Chall-Readability, and McAlpine EFLAW Readability). The results were summarized with possible explanations.

The advantage of using the cathoven analysis compared to many other methods is an unsupervised analysis, i.e. extracting the principal components, the label does not require a set of data. this is one of the requirements for an automated system classification of texts that are without human examples of new tags registered and to Add training data. While using the manual method created a high volume of the feature vector. The main components are associated with a large error and probably there are leaving features inactive or removing existing features. The fact that texts require large content systems to access information encourages us to use classification to evaluate performance it is placed and compares the basic results more reliably and easily in the future evaluation work

## **Implications**

As part of this analysis, field experts used a grid developed with traditional formulas to classify YDS reading comprehension texts. While, a modern approach was used to classify YDS reading comprehension texts using Machine Learning algorithms. A comparison between text classification scores obtained by machine learning algorithms and those obtained by field experts is the first research question and subject of the current study. Regarding the first research question, there is a statistically significant difference between traditional formulas used by field experts and modern data-based methods (Cathoven). CATHOVEN program shows higher readability than field experts. Due to traditional grids' focus on surface elements (such as syllables, words, or sentence length), metacognitive (e.g. metadiscoursal) factors were

not taken into account. As a result, traditional formulas do not have content validity when used as parameters for English difficulty/complexity.

The second research question examined the use of modern data-based methods for difficulty classification and scoring of YDS reading comprehension texts. Based on the results, CATHOVEN provided reliable answers to the questions and due to the fact that its score was 0.918 ( $P > 0.05$ ). This study improved the validity and reliability of the description of sentence complexity by incorporating additional complex features such as word frequency, word dictionary, text morphology, and the depth of the parse tree.

The third research question aimed to analyze the difficulty classification and scoring of YDS reading-comprehension texts in the measurements made by English assessment-evaluation experts over traditional formulas. As a result, for the ELT academicians who were invited to participate via email, Syntactic depth, Sentence-length, Difficult words, Vowels, and Nominal quotient (NQ) were suggested to consider to ensure the survey's validity. By analyzing the answers given by academicians to the survey's questions, the survey's reliability was calculated. Since, in traditional readability models, comprehension factors such as coherence, meaning construction, and cognitive processes such as decoding and syntactic parsing are not taken into consideration, consequently, unverifiable answers were found in the results. Therefore, field experts' answers regarding comprehension texts analysis contain more mistakes and It is evident that this type of test is prone to human error.

In contrast to the traditional formulas resulting in faulty or missing measurements in English text classification, machine learning algorithms had efficiency scores. In the area of English as a foreign language, machine learning algorithms have paved the way for development of suitable instructional materials and measuring instruments, and significant advances have been made

### **Limitations and Suggestions for Further Research**

This study had several limitations. It was primarily limited by the voluntary nature of the online responses. While many follow-up emails and reminders were sent to academicians, the participation in the online system was low. In addition, academicians who participated in the study were not all at the same academic level; some were assistant professors and others were associate professors. As well The results may have been inconsistent because some of them graduated from English-speaking universities. Moreover, the most texts should be easy for ELT faculty members as our target community is faculty members with a specialization in ELT.

Further research will be conducted in this area, university students who have taken this test or are preparing for it should be used, because the academicians may perceive all the questions as easy since they have a high level of knowledge in the field, which can result in test errors. There is also the option of using other artificial intelligence program that considers more classification parameters and also make a comparison between the results on that program and the results on Cathoven program.

## **Conclusion**

This study aimed to compare the text classification scoring made by field experts with the grid developed with traditional formulas on YDS reading-comprehension texts and the text classification scoring made using machine learning algorithms, which is one of the modern approaches over important parameters and with this scoring comparison, it is aimed to evaluate the use of machine learning algorithms in text classification according to the readability/difficulty level for measurement and evaluation. As a vital part of text mining is assessing the readability/difficulty of the text, it is important to assess the difficulty of texts in order to contribute to the teaching and learning process in education and training. It is crucial to determine the readability of comprehension texts. In this study, the readability and difficulty levels of YDS reading texts were assessed using readability formulas. As a whole, the results indicated that with the machine learning method (cathoven) by using the combined classification methods, the quality of classification improved and Cathoven classified the texts in a reliable manner while there were more errors in the answers provided by field experts regarding comprehension texts analysis.

## REFERENCES

- Abualigah, L., Gandomi, A. H., Elaziz, M. A., Hamad, H. A., Omari, M., Alshinwan, M., & Khasawneh, A. M. (2021). Advances in meta-heuristic optimization algorithms in big data text clustering. *Electronics*, 10(2), 101.
- Al Nuaimi, E., Al Neyadi, H., Mohamed, N., & Al-Jaroodi, J. (2015). Applications of big data to smart cities. *Journal of Internet Services and Applications*, 6(1), 1-15.
- Aluisio, S., Specia, L., Gasperin, C., & Scarton, C. (2010). Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 fifth workshop on innovative use of NLP for building educational applications* (pp. 1-9). <https://aclanthology.org/W10-1001.pdf>
- Ambati, B.R., Reddy, S., & Steedman M. (2016). Assessing relative sentence complexity using an incremental CCG parser. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, California, United States, pp. 1051–1057.
- Ardabili, S. F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A. R., Reuter, U., & Atkinson, P. M. (2020). Covid-19 outbreak prediction with machine learning. *Algorithms*, 13(10), 249. <https://doi.org/10.3390/a13100249>.
- Assegie, T. A. (2021). Support vector machine and k-nearest neighbor based liver disease classification model. *Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics*, 3(1), 9-14. <https://doi.org/10.35882/ijeemi.v3i1.2>
- Ateşman, E. (1997). Türkçede okunabilirliğin ölçülmesi. *Dil Dergisi*, (58),71-74.
- Azarian, R. (2011). Potentials and limitations of comparative method in social science. *International Journal of Humanities and Social Science*, 1(4), 113-125.
- Badarudeen, S., & Sabharwal, S. (2010). Assessing readability of patient education materials: current role in orthopaedics. *Clinical Orthopaedics and Related Research*, 468, 2572-2580. <https://doi.org/10.1007/s11999-010-1380-y>
- Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24, 63-88. <https://doi.org/10.1007/s10648-011-9181-8>
- Bezirci, B., & Yilmaz, A. E. (2010). Metinlerin okunabilirliğinin ölçülmesi üzerine bir yazılım kütüphanesi ve Türkçe için yeni bir okunabilirlik ölçütü. *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi*, 12(3), 49-62.
- Bhavani, A., & Kumar, B. S. (2021, April). A review of state art of text classification algorithms. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1484-1490). IEEE.
- Boateng, E. Y., Otoo, J., & Abaye, D. A. (2020). Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: a review. *Journal of Data Analysis and Information Processing*, 8(4), 341-357. <https://doi.org/10.4236/jdaip.2020.84020>
- Bonner, E., Lege, R., & Frazier, E. (2023). Large language model-based artificial intelligence in the language classroom: practical ideas for teaching. *Teaching English with Technology*, 23(1).
- Bormuth, J. R. (1968). Cloze test readability: Criterion reference scores. *Journal of Educational Measurement*, 5(3), 189-196.

- Boztaş, N., Özbilgin, Ş., Öçmen, E., Altuntaş, G., Özkardeşler, S., Hancı, V., & Günerli, A. (2014). Anestezi Uygulaması Öncesi Kullanılan Aydınlatılmış Onam Formlarının Okunabilirliklerinin Değerlendirilmesi: Karşılaştırmalı Bir Çalışma. *Turkish Journal of Anaesthesiology & Reanimation*, 42, 140-4.
- Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, 300, 70-79.
- Çetinkaya, G. (2010). Türkçe metinlerin okunabilirlik düzeylerinin tanımlanması ve sınıflandırılması, (*Yayınlanmamış Doktora Tezi*), Ankara Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.  
<https://dspace.ankara.edu.tr/xmlui/bitstream/handle/20.500.12575/33947/?sequence=1>
- Chang, C.C., & Lin, C.J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2(3), 1–27.
- Chen, J., Gong, Z., & Liu, W. (2019). A nonparametric model for online topic discovery with word embeddings. *Information Sciences*, 504, 32-47.
- Chen, J., Gong, Z., & Liu, W. (2020). A Dirichlet process biterm-based mixture model for short text stream clustering. *Applied Intelligence*, 50, 1609-1619.
- Chen, X., & Meurers, D. (2019). Linking text readability and learner proficiency using linguistic complexity feature vector distance. *Computer Assisted Language Learning*, 32(4), 418-447. <https://doi.org/10.1080/09588221.2018.1527358>
- Cheng, Y., Yao, L., Xiang, G., Zhang, G., Tang, T., & Zhong, L. (2020). Text sentiment orientation analysis based on multi-channel CNN and bidirectional GRU with attention mechanism. *IEEE Access*, 8, 134964-134975.  
<https://doi.org/10.1109/ACCESS.2020.3005823>
- Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2), 97-135. <https://doi.org/10.1075/itl.165.2.01col>
- Crossley, S. A., Allen, D. B., & McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language*, 23(1), 84-101. <https://eric.ed.gov/?id=EJ926371>
- Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational Research Bulletin*, 37-54. <https://www.jstor.org/stable/1473669>
- Dale, E., & Chall, J.S. (1949)“The concept of readability,” *Elementary English*, 26(1), 19–26..
- Danielson, K.E. 1987. Readability formulas: a necessary evil? *Reading Horizons: A Journal of Literacy and Language Arts* 27(3):4.
- Dawson, B., & Trapp, R.G. (2001). *Probability&related topics for making inferences about data*. Lange medical Books/McGraw-Hill Medical Publishing Division.
- Debortoli, S., Müller, O., Junglas, I., & Vom Brocke, J. (2016). Text mining for information systems researchers: An annotated topic modeling tutorial. *Communications of the Association for Information Systems (CAIS)*, 39(1), 7.  
<https://doi.org/10.17705/1CAIS.03907>
- Deepa, M. D. (2021). Bidirectional encoder representations from transformers (BERT) language model for sentiment analysis task. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(7), 1708-1721.

- Devi, S. A., & Kumar, S. S. (2020). A hybrid document features extraction with clustering based classification framework on large document sets. *International Journal of Advanced Computer Science and Applications*, 11(7).
- DOI 10.1109/ACCESS.2020.2972591.
- Dong, L., Du, H., Mao, F., Han, N., Li, X., Zhou, G., & Liu, T. (2019). Very high resolution remote sensing imagery classification using a fusion of random forest and deep learning technique—Subtropical area for example. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 113-128. <https://doi.org/10.1109/JSTARS.2019.2953234>
- DuBay, W. H. (2004). The principles of readability. *Online Submission*. P: 76. <https://eric.ed.gov/?id=ed490073>
- El Naqa, I., & Murphy, M. J. (2015). What is machine learning? (pp. 3-11). *Springer International Publishing*. [https://doi.org/10.1007/978-3-319-18305-3\\_1](https://doi.org/10.1007/978-3-319-18305-3_1)
- Eom, J. H., & Zhang, B. T. (2004). PubMiner: machine learning-based text mining for biomedical information analysis. *Genomics & Informatics*, 2(2), 99-106. <https://koreascience.kr/article/JAKO200407921776300.pdf>
- Ethridge, D.E. (2004) *Research methodology in Applied Economics*. John Wiley & Sons
- Fanguy, R., Kleen, B., & Soule, L. (2004). Privacy policies: cloze test reveals readability concerns. *Issues in Information Systems*, 5(1), 117-123.
- Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010). A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics*, Stroudsburg, PA, pp. 276–284.
- Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3), 221. <https://doi.org/10.1037/h0057532>
- Fox, W., & Bayat, M.S. (2007). *A Guide to managing research*. Juta Publications
- Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education (8th ed.)*. New York, USA: McGraw-Hill.
- François, T., & Fairon, C. (2012). An “AI readability” formula for French as a foreign language. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 466-477). <https://aclanthology.org/D12-1043.pdf>
- François, T., & Miltsakaki, E. (2012). Do NLP and machine learning improve traditional readability formulas? In *Proceedings of the First Workshop on Predicting and Improving Text Readability for Target Reader Populations, Association for Computational Linguistics*, Stroudsburg, PA, pp. 49–57.
- Franks, P. C. (2013). *Records and information management*. American Library Association.
- Freda, M. C. (2005). The readability of American Academy of Pediatrics patient education brochures. *Journal of Pediatric Health Care*, 19(3), 151-156. <https://doi.org/10.1016/j.pedhc.2005.01.013>
- Fredrickson, G. M. (1997). *The comparative imagination: On the history of racism, nationalism, and social movements*. California: Univ of California Press.
- Fry, E. (1965). *Teaching faster reading*. Cambridge University Press.

- Fry, E. (1968). A readability formula that saves time. *Journal of Reading*, 11(7), 513-578. <https://www.jstor.org/stable/40013635>
- Fry, E. (2002). Readability versus leveling. *The Reading Teacher*, 56(3), 286-291. <https://www.jstor.org/stable/20205195>
- Gay, L. R., Mills, G. E., & Airasian, P. W. (2012). *Educational research: Competencies for analysis and applications (10th ed.)*. Upper Saddle River: Pearson.
- Gernsbacher, M. A. (1997). Two decades of structure building. *Discourse Processes*, 23(3), 265-304. <https://doi.org/10.1080/01638539709544994>
- Gil Ortego, R., & Martínez Sánchez, I. (2019). "Relevant Parameters for the Classification of Reading Books Depending on the Degree of Textual Readability in Primary and Compulsory Secondary Education (CSE) Students," in *IEEE Access*, vol. 7, pp. 79044-79055, doi: 10.1109/ACCESS.2019.2922608.
- Gilliam, B., Peña, S. C., & Mountain, L. (1980). The Fry graph applied to Spanish readability. *The Reading Teacher*, 33(4), 426-430.
- Graesser, A.C., McNamara, D.S., & Kulikowich, J.M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5), 223-234.
- Graesser, A.C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, 101(3), 371-395.
- Greene, B. (2001). Testing reading comprehension of theoretical discourse with cloze. *Journal of Research in Reading*, 24(1), 82-98. <https://doi.org/10.1111/1467-9817.00134>
- Gunning, R. (1952). *The technique of clear writing*. New York: McGraw-Hill.
- Guthrie, J. T. (2007) *Engaging adolescents in reading*. Thousand Oaks, CA, USA: Corwin Press.
- Güyer, T., Temur, T., & Solmaz, E. (2009). Bilgisayar destekli metin okunabilirliği analizi. *Türk Eğitim Bilimleri Dergisi*, 7(4), 751-766. <https://dergipark.org.tr/en/download/article-file/256290>
- Hancke, G. P., de Carvalho e Silva, B., & Hancke Jr, G. P. (2012). The role of advanced sensing in smart cities. *Sensors*, 13(1), 393-425. <https://doi.org/10.3390/s130100393>
- Hancke, J., Vajjala, S., & Meurers, D., (2012). Readability classification for German using lexical, syntactic, and morphological Features. *Proceedings of COLING 2012: Technical Papers*, 1063 – 1080.
- Hirschfeld, L.A., & Gelman, S.A. (1994). *Mapping the mind: Domain-specificity in cognition and culture*. Cambridge University Press.
- Howcroft, D.M., & Demberg, V. (2017). Psycholinguistic models of sentence processing improve sentence readability rank- ing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Valencia, Spain, pp. 958-968.
- Huerta, J. F. (1958). Legibilidad y lecturabilidad: Dos conceptos básicos en los libros escolares, *Consigna (Revista pedagógica de la sección femenina de Falange ET y de las JONS*, vol. 213, p. 3.
- Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.
- Kandula, S., & Zeng-Treitler, Q. (2008). Creating a gold standard for the readability measurement of health texts. In *AMIA Annual Symposium Proceedings (Vol. 2008, p.*

- Kanebrant, E., Mühlenbock, K.H., Kokkinakis, S.J., Jönsson, A., Liberg, C., Geijerstam, Å., Folkeryd, J.W., & Falkenjack, J. (2015). T-MASTER – A tool for assessing students' reading abilities. In *Proceedings of the 7th International Conference on Computer Supported Education (CSEDU 2015)*, Lisbon, Portugal, pp. 220–227.
- Kanungo, T., & Orr, D. (2009). Predicting the readability of short web summaries. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining* (pp. 202-211). <https://doi.org/10.1145/1498759.1498827>
- Karatay, H., Bolat, K. K., & Güngör, H. (2013). Türkçe ders kitaplarındaki metinlerin okunabilirlik ve anlaşılabilirliği. *The Journal of Academic Social Science Studies*, 6(6), 603-623. <http://dx.doi.org/10.9761/JASSS1770>
- Kesavan, N., Jaisankar, N., & Ramani, S. (2017). Pattern taxonomy mining for text categorization. *International Journal of Computer Applications*, 168(8).
- Keshavarz, M. H., & Salimi, H. (2007). Abstract. *International Journal of Applied Linguistics*, 17(1), 81-92.
- Kintsch, W., & Vipond, D. (2014). 'Reading comprehension and readability in educational practice and psychological theory, in *Perspectives on Learning and Memory*, L.-G. Nilsson, Ed. Hillsdale, NJ, USA: Erlbaum, pp. 329–365.
- Klare, G.R. (1963). *The Measurement of Readability*. Ames, IA: Iowa State University Press.
- Klare, G.R. (2000). The measurement of readability: Useful information for communicators. *ACM Journal of Computer Documentation (JCD)* 24(3), 107–121.
- Kou, G., Yang, P., Peng, Y., Xiao, F., Chen, Y., & Alsaadi, F. E. (2020). Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. *Applied Soft Computing*, 86, 105836.
- Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, 10(4), 150.
- Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, 10(4), 150.
- Kumbhar, P., & Mali, M. (2016). A survey on feature selection techniques and classification algorithms for efficient text classification. *International Journal of Science and Research*, 5(5), 9.
- Kumiko, T., Satoshi, T., & Hirosh, T. (2010). Sorting texts by readability. *Association for Computational Linguistics*, 36(2), 203 – 228.
- Larose, D. T., & Larose, C. D. (2014). *Discovering knowledge in data: an introduction to data mining* (Vol. 4). John Wiley & Sons.
- Lee, J. Y., & Dernoncourt, F. (2016). Sequential short-text classification with recurrent and convolutional neural networks. *arXiv preprint arXiv:1603.03827*.
- Liu, C., & Wang, X. (2020). Quality-related English text classification based on recurrent neural network. *Journal of Visual Communication and Image Representation*, 71, 102724.
- Liu, P., Qiu, X., Chen, X., Wu, S., & Huang, X. J. (2015). Multi-timescale long short-term memory neural network for modelling sentences and documents. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 2326-2335).

- Luo, X. (2021). Efficient English text classification using selected machine learning techniques. *Alexandria Engineering Journal*, 60(3), 3401-3409.
- Ly, L., & Liu, Y. S. (2005, December). Research of english text classification methods based on semantic meaning. In *2005 international conference on information and communication technology* (pp. 689-700). IEEE.
- Macahilig, H.B. (2014). A Content-Based Readability Formula for Filipino Texts. *The Normal Lights: Journal on Teacher Education*, 8(1), ICEAP-17 .
- Maqsood, S., Shahid, A., Nazar, F., Asif, M., Ahmad, M., Mazzara, M. (2020). C-POS: A context-aware adaptive part-of-speech language learning framework. *IEEE Access* 8, 30 720 –30 733
- Marsh, K. (2022). Comparing Text Type and Inferences (*Doctoral dissertation, Lesley University*).
- Mayfield, E., & Penstein-Rosé, C. (2010). Using feature construction to avoid large feature spaces in text classification. In *Proceedings of the 12th annual conference on Genetic and evolutionary computation* (pp. 1299-1306).
- Mc Laughlin, G. H. (1969). SMOG grading-a new readability formula. *Journal of Reading*, 12(8), 639-646. <https://www.jstor.org/stable/40011226>
- McConnell, C.R. (1982). Readability formulas as applied to college economics textbooks. *Journal of Reading* 26(1), 14–17.
- McGilvray, D. (2021). *Executing data quality projects: Ten steps to quality data and trusted information (TM)*. Academic Press.
- McMillan, J., & Schumacher, S. (2014). *Research in education evidence-based Inquiry (7th ed.)*. the United State of America: Pearson.
- McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension. *Psychology of Learning and Motivation*, 51, 297-384. [https://doi.org/10.1016/S0079-7421\(09\)51009-2](https://doi.org/10.1016/S0079-7421(09)51009-2)
- McNamara, D.S., Louwerse, M.M., & Graesser, A.C. (2002). *Coh-Metrix: Automated cohesion and coherence scores topredict text readability and facilitate comprehension*. Technical report, Institute for Intelligent Systems, University of Memphis, Memphis, TN.
- McNamara, D.S., Louwerse, M.M., McCarthy, P.M., & Graesser, A.C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*, 47(4), 292–330.
- Miltsakaki, E., & Truett, A. (2007). Read-x: Automatic evaluation of reading difficulty of web text. In *Proceeding of the World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, Vol. 2007, No. 1, Quebec, Canada, pp. 7280–7286.
- Mirzaoglu, V., & Erhan, A. (2015). 5. Sınıf türkçe ders kitabındaki metinlerin okunabilirliği üzerine bir inceleme. *Siirt Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 5, 146-155. <https://dergipark.org.tr/en/pub/susbid/issue/17332/180997>
- Moreno, A., & Redondo, T. (2016). Text analytics: the convergence of big data and artificial intelligence. *IJIMAI*, 3(6), 57-64.
- Okur, A., & Arı, G. (2013). 6, 7, 8. sınıf Türkçe ders kitaplarındaki metinlerin okunabilirliği. *İlköğretim Online*, 12(1), 202-226.
- Ott, B. B., & Hardie, T. L. (1997). Readability of advance directive documents. *Image: The Journal of Nursing Scholarship*, 29(1), 53-57. <https://doi.org/10.1111/j.1547-5069.1997.tb01140.x>

- Parente, D. J. (2021). PolyBoost: an enhanced genomic variant classifier using extreme gradient boosting. *PROTEOMICS–Clinical Applications*, 15(2-3), 1900124. <https://doi.org/10.1002/prca.201900124>
- Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The acquisition of Reading Comprehension Skill. In M. J. Snowling & C. Hulme (Eds.), *The science of Reading: A Handbook* (pp. 227–247). <https://doi.org/10.1002/9780470757642.ch13>
- Petersen, S.E., & Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer Speech & Language*, 23(1), 89–106.
- Pickvance, C. (2005). The four varieties of comparative analysis: the case of environmental regulation. *Journal of Housing and the Built Environment*, 16, 7-28.
- Pop, D. (2016). Machine learning and cloud computing: Survey of distributed and saas solutions. *arXiv preprint arXiv*, 1603.08767. <https://doi.org/10.48550/arXiv.1603.08767>
- Rasane, S., & Patil, D. V. (2016). Handling various issues in text classification: a review. *International Journal on Emerging Trends in Technology*, 3(1), 4076-4082.
- Reddy, D. L. C. (2011). A review on data mining from past to the future. *International Journal of Computer Applications*, 975, 8887.
- Richards, J. C., & Schmidt, R. (2010). *Dictionary of language teaching & applied linguistics* (4th ed.). Harlow, Great Britain: Pearson Education Limited.
- Richardson, J. S., Morgan, R. F., & Fleener, C. (2012). Reading to learn in the content areas. *Cengage Learning*. 504 pages
- Rostami, M., Ayat, S. S., Attarzadeh, I., & Saghari, F. (2015). Proposing a method to classify texts using data mining. *Journal of Advances in Computer Research*, 6(4), 125-137.
- Rousseau, F., Kiagias, E., & Vazirgiannis, M. (2015). Text categorization as a graph classification problem. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers) (pp. 1702-1712).
- Routman, R. (2003). *Reading Essentials: The Specifics You Need to Teach Reading Well*. London, U.K.: Pearson
- Rubenstein, H., & Aborn, M. (1958). Learning, prediction, and readability. *Journal of Applied Psychology*, 42(1), 28–32.
- Rufo, D. D., Debelee, T. G., Ibenthal, A., & Negera, W. G. (2021). Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM). *Diagnostics*, 11(9), 1714. <https://doi.org/10.3390/diagnostics11091714>
- Rush, RT. 1985. Assessing readability: formulas and alternatives. *The Reading Teacher*, 39(3), 274–283.
- Savla, P., & Martino, L. D. (2012). Content analysis of privacy policies for health social networks. In *2012 IEEE International Symposium on Policies for Distributed Systems and Networks* (pp. 94-101). IEEE. <https://doi.org/10.1109/POLICY.2012.20>
- Si, L., & Callan, J. (2001) A statistical model for scientific readability, in *Proc. 10th Int. Conf. Inf. Knowl. Manage. (CIKM)*, pp. 574–576.
- Snow, C. E (2002). *Reading for understanding: Toward an r&d program in reading comprehension*. Santa Monica, CA, USA: RAND Corporation.

- Solmaz, E. (2009). İlköğretim 4. ve 5. Sınıf Düzeylerindeki Türkçe Metinlerde Cümle Uzunluğu, Kelime Uzunluğu, Kelime Uzunluğu ve Kelime Uzunluğu ve Kelime Hazinesinin Okunabilirlik Üzerine Etkisi. (*Yüksek Lisans Tezi*), Gazi Üniversitesi.
- Stevens, E.A., Walker, M. A., & Vaughn, S. (2017). The effects of reading fluency interventions on the reading fluency and reading comprehension performance of elementary students with learning disabilities: A synthesis of the research from 2001 to 2014, *Journal of Learning Disabilities*, 50(5), 576–590.
- Stobart, G., & Eggen Theo, J.H.M. (2012). High-stakes testing – value, fairness and consequences. *Assessment in Education Principles Policy and Practice*, 19(1), 1-6, <http://dx.doi.org/10.1080/0969594X.2012.639191>
- Stockmeyer, N. O. (2009). Using Microsoft Word's readability program. *Michigan Bar Journal*, 88, 46. <https://ssrn.com/abstract=1210962>
- Sung, Y.T., Chen, J.L., Cha, J.H., Tseng, H.C., Chang, T.H., & Chang, K.E. (2015). Constructing and validating readability models: The method of integrating multilevel linguistic features with machine learning. *Behavior Research Methods*, 47(2), 340–354.
- Sung, Y.T., Chen, J.L., Cha, J.H., Tseng, H.C., Chang, T.H., & Chang, K.E. (2015a). Constructing and validating readability models: The method of integrating multilevel linguistic features with machine learning. *Behavior Research Methods*, 47(2), 340–354.
- Tai, K. S., Socher, R., & Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.
- Tekşan, K. (2016). Ömer Seyfettin'in Hikâyelerinin Okunabilirlik Düzeyleri. *Sosyal ve Liberal Bilimlerde Yeni Yönelimler*, 1, 25-36.
- Thorndike, E.L., & Lorge, I. (1944). *The Teacher's Word Book of 30,000 Words*. New York: Teachers College, Columbia University. Bureau of Publications.
- Tilly, C. (1984). *Big structures, large processes, huge comparisons*. Russell Sage Foundation.
- Tseng, H. C., Chen, B., Chang, T. H., & Sung, Y. T. (2019). Integrating LSA-based hierarchical conceptual space and machine learning methods for leveling the readability of domain-specific texts. *Natural Language Engineering*, 25(3), 331-361.
- Tseng, H.C., Chen, B., Chang, T.H., & Sung, Y.T. (2019). Integrating LSA-based hierarchical conceptual space and machine learning methods for leveling the readability of domain-specific texts. *Natural Language Engineering*, 25(3), 331–361. <https://doi.org/10.1017/S1351324919000093>
- Vajjala, S., & Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications using NLP*, Montreal, Canada, pp. 163–173.
- Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, 5(1), 7-16.
- Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, 5(1), 7-16.
- Wissing, G. J., Blignaut, A. S., & Van den Berg, K. (2016). Using readability, comprehensibility and lexical coverage to evaluate the suitability of an introductory accountancy textbook to its readership. *Stellenbosch Papers in Linguistics*, 46, 155-179. <https://doi.org/10.5774/46-0-205>

- Woods, B., Moscardo, G., & Greenwood, T. (1998). A critical review of readability and comprehensibility tests. *Journal of Tourism Studies*, 9(2), 49-61. <https://search.informit.org/doi/abs/10.3316/ielapa.990807829>
- Yan, X., Song, D., & Li, X. (2006). Concept-based document readability in domain-specific information retrieval. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, ACM, New York, NY, pp. 540–549.
- Yuan, J., Jin, Z., Guo, H., Jin, H., Zhang, X., Smith, T., & Luo, J. (2020). Constructing biomedical domain-specific knowledge graph with minimum supervision. *Knowledge and Information Systems*, 62, 317-336. <https://doi.org/10.1007/s10115-019-01351-4>
- Zhang, J., Chang, W. C., Yu, H. F., & Dhillon, I. (2021). Fast multi-resolution transformer fine-tuning for extreme multi-label text classification. *Advances in Neural Information Processing Systems*, 34, 7267-7280.
- Zhu, X., Sobihani, P., & Guo, H. (2015). Long short-term memory over recursive structures. In *International Conference on Machine Learning* (pp. 1604-1612). PMLR.
- Zorbaz, K. Z. (2007). An evaluation on the word-sentence lengths and readability levels of tales in Turkish textbooks. *Journal of Theory and Practice in Education*, 3(1), 87-101.

## APPENDICES

### APPENDIX-1. KATILIMCI RIZA FORMU VE BİLGİLENDİRME METNİ

Değerli Katılımcı,

Öncelikle yapacağım bu çalışmaya gösterdiğiniz ilgi ve bana ayırdığınız zaman için teşekkür ederim. Bu form, araştırmanın amacını ve sizin bir katılımcı olarak haklarınızı tanımlamayı amaçlamaktadır.

Bu çalışmada, YDS okuma-anlama metinleri üzerinde geleneksel formüller ile geliştirilmiş grid ile alan uzmanları tarafından yapılan metin sınıflandırma puanlamaları ile modern yaklaşımlardan olan makine öğrenim algoritmaları kullanılarak yapılan metin sınıflandırma puanlamalarının önemli parametreler üzerinde karşılaştırmak ve bu puanlama karşılaştırması ile makine öğrenim algoritmalarının ölçme değerlendirme için okunurluk/zorluk düzeyine göre metin sınıflandırmadaki kullanımın değerlendirilmesi amaçlanmaktadır.

Araştırmaya gönüllü olarak katılmanızın ve ifade edeceğiniz görüşlerinizin çalışmaya ışık tutacağına ve büyük katkıları olacağına inanıyorum. Size özel bilgileri açıklamanızda herhangi bir zorlama olmayacaktır. Kayda alınacak bu görüşme, yalnızca bilimsel bir veri olarak bu araştırma için kullanılacak ve bunun dışında hiçbir amaçla kullanılmayacaktır. Veriler sadece araştırmacının ulaşabileceği güvenli bir ortamda saklı tutulacaktır. İsterseniz yazılı dokümanlar incelemeniz üzere size verilecektir. Çalışmanın yazılı hale getirilmesinde yapılacak tüm alıntılar için gerçek isminizin yerine takma bir isim (Örn: K1, Ö1 gibi) kullanılacak; sizi deşifre edecek her türlü unsur, değişik bir isimle ya da örtük olarak ifade edilecektir. İstedığınız zaman anketi bitirebilir ve çalışmadan ayrılabilenizi belirtmek isterim. Böyle bir durumda yazılan raporları size teslim edeceğim. Dilerseniz görüşme yaptıktan sonra da araştırmadan çekilebilirsiniz. Bu durumda bütün kayıtlar size teslim edilecek ve araştırmada kesinlikle kullanılmayacaktır.

Bu sözleşmeyi okuyup, bu araştırmaya gönüllü olarak katıldığınıza ve araştırma kapsamında benim size verdiğim güvenceye ilişkin olarak bu formu imzalamanızı rica ediyorum. Bu sözleşmeyi okuyarak kabul kutucuğunu işaretlediğini için teşekkür ederim. Ayrıca, alta yer alan iletişim bilgileri ile araştırmacı ve tez danışmanı ile araştırmayla ilgili öneri, soru ve eleştirileriniz için her zaman iletişime geçebilirsiniz.

Araştırmacı: Somayyeh SHABESTANI

Atatürk Üniversitesi Yabancı Diller Eğitimi Bölümü İngiliz Dili ve Eğitimi A.B.D.

Tel: 0(539) 510 11 02

e-mail: [somayyeh.shabestani20@ogr.atauni.edu.tr](mailto:somayyeh.shabestani20@ogr.atauni.edu.tr)

Tez Danışmanı: Doç. Dr. Merve GEÇİKLİ

Atatürk Üniversitesi Yabancı Diller Eğitimi Bölümü İngiliz Dili ve Eğitimi A.B.D.

Tel: 0(545) 877 57 61 – 0(442) 231 44 92

e-mail: [merve.gecikli@atauni.edu.tr](mailto:merve.gecikli@atauni.edu.tr) , [merve.gecikli25@gmail.com](mailto:merve.gecikli25@gmail.com)

## APPENDIX-2. GRID

Değerli Öğretim Üyesi,

Bu grid, geleneksel formüller üzerinden YDS sınavlarındaki okuma-anlama metinlerinin okunabilirlik bakımından zorluk seviyelerinin sentaktik, semantik, anlamsal ve topolojik parametrelerden hareketle sınıflandırılmasını amaçlamaktadır. Bu nedenle tüm sorulara cevap vermeniz oldukça önemlidir. Vereceğiniz cevaplar sadece araştırma amacıyla “ Makina Öğrenim Algoritmaları İle İngilizce Metin Zorluk Seviyelerinin Sınıflandırılması: Bir Yazılım Önerisi” adlı yüksek lisans tezi kapsamında kullanılacaktır. Grid’de yapacağınız puanlamalar makine öğrenim algoritmaları ölçümleri ile karşılaştırılarak ölçme-değerlendirmede seviyeye uygun metin seçiminde metin sınıflandırma sisteminde makine öğrenim algortimalarının etkinliğini değerlendirmede kullanılacaktır.

Griddeki metinleri dikkatlice okuduktan sonra lütfen seçeneklerden size en uygun olanını boşluklara (x) işareti koyarak yanıtlayınız. Araştırmaya yapacağınız katkılardan dolayı sizlere teşekkür ederiz.

### A.KİŞİSEL BİLGİLER

#### 1.Cinsiyetiniz

1. ( ) Kadın

2. ( ) Erkek

#### 2.. Mesleki Kıdeminiz

1.( ) 0-5 yıl

2.( ) 6-10 yıl

3.( ) 11-15 yıl

4. ( ) 16-20 yıl

5.( ) 20 ve üzeri  
yıl

### B. METİN OKUNURLUK PUANLAMASI

#### METİN 1

Viral infections of the respiratory tract are certainly the most common cause of infectious illness in most countries. People in all age groups are susceptible, but for those at the extremes of life these illnesses are particularly hazardous. Viral respiratory infections are more common in children under the age of five years, but in the first year of life, they are more severe. In the elderly, degenerative processes of the heart or lungs make pneumonia a more frequent and serious complication, and the same is true for patients of all ages suffering from chronic cardiac disease or chronic bronchitis. On the whole, the vast majority of viral respiratory infections are mild though often uncomfortable conditions, but sometimes the illness is severe and constitutes a threat to life. Unfortunately, there is no method whereby the spread of infection can be prevented. Adequate ventilation and avoidance of crowds are clearly advisable.

#### ZORLUK SEVİYESİ (SENTANTİK)

KOLAY

ORTA

ZOR

1

2

3

4

5

6

7

8

9

#### ZORLUK SEVİYESİ (SEMANTİK)

|                             |   |   |      |   |   |     |   |   |
|-----------------------------|---|---|------|---|---|-----|---|---|
| KOLAY                       |   |   | ORTA |   |   | ZOR |   |   |
| 1                           | 2 | 3 | 4    | 5 | 6 | 7   | 8 | 9 |
| ZORLUK SEVİYESİ (SÖZLÜKSEL) |   |   |      |   |   |     |   |   |
| KOLAY                       |   |   | ORTA |   |   | ZOR |   |   |
| 1                           | 2 | 3 | 4    | 5 | 6 | 7   | 8 | 9 |
| ZORLUK SEVİYESİ (TOPOLOJİK) |   |   |      |   |   |     |   |   |
| KOLAY                       |   |   | ORTA |   |   | ZOR |   |   |
| 1                           | 2 | 3 | 4    | 5 | 6 | 7   | 8 | 9 |



### APPENDIX-3. YDS CORPUS

1. Globalization is reflected in certain areas of life such as culture and lifestyle. Modern mass media and increased mobility favor a sort of cultural globalization. African cooking and Indian films have become as common in Europe as Western fast food is in Asia or Hollywood films in the Middle East. Optimists see this mingling of world societies as a chance to integrate 'the foreign' into one's own cultural value system and in this way to increase mutual tolerance. Growing commonalities in the sense of a recognized universal value system, such as human rights, can develop in this way. This perspective presupposes free access to information and knowledge. In contrast, critics emphasize the economic dominance of rich industrial nations in the media, through which they force their Western model of affluence on the weaker countries for their own economic advantage. The general commercialization and reshaping of national or regional cultures through foreign influences have, in many parts of the world, provoked mass movements seeking a return to their own traditional values. One can trace the radical anti-Western movements back to these perceived causes. The emphasis on regional, local, and new nationalist thinking can partly be seen as a reaction to globalization (YDS, 2013).
2. For fifty years, scientists have been trying to create tiny gadgets that travel around a patient's bloodstream, detecting and even fixing diseases. But how to power them has been a problem. Formerly-used batteries have shrunk over the years but have not been miniaturized sufficiently to fit into a medical device small enough to travel through veins and arteries. So instead of using them, engineers at Stanford University have developed tiny implants that can be powered by radio waves beamed from a transmitter outside the body. The idea of using electromagnetic waves to transfer power to implantable medical devices is not new, and in principle, it sounds simple. A transmitter outside the body emits radio waves travelling through the tissues. They are then picked up by an antenna on the internal device where they induce a voltage. The trouble is, scientists previously thought that radio waves would struggle to get through human tissue. But a researcher at Stanford University found strong evidence that they do pass through. He used a chunk of beef as a stand-in for human flesh and found that sufficient energy moved through it to make electromagnetic power transfer feasible. Already, working prototypes that can propel themselves have been developed, but the distance through which power can be transferred inside the body still needs improvement (YDS, 2014).
3. Growing orchids could change your life - forever! It has for many people. No other plant family offers the number and diversity of breathtaking and intriguing flowers that orchids do. When you discover that you can actually grow these exotic plants in your home and that they are not expensive, your resistance will be lowered and your chances of catching the extremely contagious but wonderful obsession called the 'Orchid Bug' are high. In England, and the rest of Europe in Victorian times, these fabulous plants used to be only for the eyes of royalty or well-heeled aristocrats because of the one-time astronomical prices they commanded at flower auctions. They were grown in elaborate

and costly glasshouses, usually by a staff of professional growers. However, as times have changed, orchids have entered the mainstream. They are now the second most popular pot plant in the world and are gaining quickly on the leader (poinsettia)! Because of new techniques and modern growing methods, award-quality orchids are available to everyone at prices lower than ever before. You can buy an orchid plant in full bloom for less than you would pay for a bouquet of flowers from the florist. In addition to the orchids of today being less expensive, they are also much easier to grow. Some bloom more than once a year and others have a pleasant smell. And with thousands of hybrids and new ones being produced all the time, there is an orchid that will strike anyone's and everyone's fancy (YDS, 2015).

4. The floods causing damage across much of central Europe are a sign of catastrophes to strike as the continent's climate gets harsher. In the German town of Passau on 3 June 2013, waters rose to their highest level since 1501. At that time, the floods in Czech capital Prague were beginning to recede but Dresden, Germany, was preparing itself for the river Elbe to rise 5 meters higher than normal. "Several factors are responsible", says Stephan Isoard of the European Environment Agency in Copenhagen, Denmark. "It was spring, so snow was melting from the mountains," she says. When two months of rain fell in two days, the water had nowhere to go because the ground was soaked. Climate change also causes heavier rainfall, and might be partly to blame. However, Isoard points out that bad land management is just as important. "In urban areas, there is less opportunity for water to infiltrate the soil. With more floods inevitable, Europe needs to adapt," Isoard says. Some work is already under way. Wetlands are being restored around stretches of the Danube. Green spaces like this can absorb extra water, making floods less severe. "Over the last 20 years, events like this have become more common," says Iain White of the University of Manchester in the UK. "Central Europe has improved its flood responses since 2002, but there comes a point where you can't defend," he maintains (YDS, 2016S).
5. One day in 1952, John W. Hetrick was driving with his wife and daughter in the front seat when he had to change his direction suddenly and brake quickly to avoid an obstacle. Instinctively, he and his wife put their arms out to protect their daughter in case of a crash. This event inspired him to provide automobiles with airbags to protect people during accidents. Hetrick had been an engineer in the US Navy during World War II. He recalled a compressed-air torpedo accidentally turning itself on, causing its canvas cover to shoot up into the air, 'quicker than a blink of an eye'. In 1952, Hetrick proposed using compressed air to inflate airbags rapidly during car crashes. He received a patent for this invention in 1953, but car manufacturers in the 1950s were more interested in style than safety. Later, consumers became more safety conscious and airbag technology improved. The first airbags were optional, but by the 1990s, they had become standard. Although airbags have saved thousands of lives, they are not always sufficient to prevent death and injury during crashes. Travelers must also wear seat belts, and automobiles must have dashboards made of soft materials and steering columns that can absorb energy during impact from car crashes (YDS, 2017S).

6. Admired by native peoples for thousands of years, the Texas horned lizard has an array of abilities. It mostly eats ants – lots of them, and most parts of an ant’s body are indigestible, which necessitates a huge stomach. Eating more than two hundred ants per day means exposure out in the open for long periods, and having a ‘heavy’ stomach means that a horned lizard finds it difficult to run away from predators. Instead, it relies on an armory of defenses. It has camouflage coloring, with an outline broken up by spines and outgrowths, and it will freeze if a predator approaches. Its horns and spines can pierce the throat of a snake or bird, and it can hiss and blow itself up to look even more fearsome. When it comes to coyotes, foxes, and dogs, a horned lizard’s most spectacular defense is to squirt foul-tasting blood from the sinuses behind its eyes. That usually has the desired effect. But it squirts only when it is provoked, since it risks losing up to one quarter of its blood. Such abilities are, however, no defense against human invasion of its land. Its strange shape and coloring has made it attractive to reptile collectors, and its habit of freezing means that it is prone to being run over (YDS, 2018S).
7. We are born to judge others by how they look: our brains come hardwired with a specific face-processing area, and even shortly after birth, babies would rather look at a human face than anything else. Within their first year, they become more discerning, and are more likely to attend to friendly-looking faces than those who look serious. By the time we reach adulthood, we develop a great number of stereotypes and become snap-judgement specialists, jumping to conclusions about a person’s character and status after seeing his or her face for just a tenth of a second. And we ignore considered assessments of others in favor of simple cognitive shortcuts. For example, we judge a baby-faced individual as more trustworthy, associate a chiseled jaw with dominance, or refer to a person with a big nose as a curious one. Unfair or unethical, it may be, but it makes good evolutionary sense. Ours is an ultra-social species, so being able to quickly assess whether someone is friend or foe and whether they have the power to help or hurt us is important survival information. But there is a problem. As psychologist Alexander Todorova of Princeton University points out, more often than not, our first impressions are wrong; that is, relying on our shortcuts may not always produce good results. It is not clear why, but he suggests that we meet many more strangers than our prehistoric ancestors would have, and this may play a role (YDS, 2019S).
8. Most people, from those who are fond of living alone to the rich and famous, probably have experienced feelings of shyness at various times in their lives. Physiological symptoms may include blushing, increased heart rate, sweating, and shaking. Just as these outward manifestations vary in type and intensity from person to person, so do the inner feelings. Anxious thoughts and worries, low self-esteem, self-criticism, and concern over a lack of social skills, real or imagined, are common. Nevertheless, the causes of shyness are not known. Some researchers believe it results from a genetic predisposition. Others theorize that uncommunicative parents restrict a child’s development of the social skills that compensate for discomfort caused by new experiences and people, resulting in shyness. Various, it has been considered a symptom of social phobia or a simple characteristic of introversion. Psychological research that follows large numbers of children from very early childhood to adulthood has found that a tendency to be shy with others is one of the most stable traits that is preserved from the first three or four years of life through young adulthood. Learning or improving social skills through self-help courses or formal training in assertiveness

and public speaking are some of the methods used to diminish the effects of shyness (YDS, 2020F).

9. So far, only two types of mental exertion have been shown to improve or preserve overall cognitive ability. One is musical training. Experts state that listening to and making music is not only an auditory experience; it is a multisensory and motor experience. Playing an instrument changes how the brain interprets and integrates a wide range of sensory information, and making music over a long period of time can change brain function and brain structure. The other is to learn a new language. As for the role of language-learning in brain power, a study carried out among students revealed that attention gets better when a new language is studied intensively even if it is for a short time and it gets even better when more practice is done for longer. The results of the study confirm the cognitive benefits of language-learning and give important and warming messages: firstly, even a short, intensive course can show beneficial effects on some cognitive functions. Secondly, this effect can be maintained through practice. So, we can say that it is never too late to start a mental activity such as learning a new language (YDS, 2021F).

## APPENDIX-4. ETİK KURUL RAPORU

T.C.  
ATATÜRK ÜNİVERSİTESİ  
SOSYAL VE BEŞERİ BİLİMLER ETİK KURULU  
Eğitim Bilimleri Birim Etik Kurul Başkanlığı  
PROJE DEĞERLENDİRME FORMU

**Projenin/Çalışmanın Adı:** Danışmanlığını Doç. Dr. Merve GEÇİKLİ'nin yürüttüğü yüksek lisans öğrencisi Somayyeh SHABESTANI'nın "Makina Öğrenim Algoritmaları İle İngilizce Metin Zorluk Seviyelerinin Sınıflandırılması: Bir Yazılım Önerisi" isimli yüksek lisans tez çalışması ile ilgili Etik Kurul uygunluk onay belgesi talep edilmesi.

Gönderildiği Tarih: 25.05.2023

Teslim Tarihi: 25.05.2023

Değerlendirenin Adı Soyadı: Sosyal ve Beşeri Bilimler Etik Kurulu Başkanlığı Adına

**Prof. Dr. Ufuk ŞİMŞEK**  
Etik Kurul Başkanı

İmza :

|                                                                                                                                                                                                       |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Bu Yüksek Lisans Tez uygundur:</b> <input checked="" type="checkbox"/>                                                                                                                             |
| <b>Gerekçe/Açıklama:</b> Yapılacak çalışmada yaklaşım ve yöntemler dikkate alınarak, konu ile ilgili çalışmaların gerçekleştirilmesinde etik ve bilimsel açıdan herhangi bir sakınca bulunmamaktadır. |
| <b>Bu Yüksek Lisans Tez uygun değildir:</b> <input type="checkbox"/><br><b>Gerekçe/Açıklama</b>                                                                                                       |

T.C.  
ATATÜRK ÜNİVERSİTESİ REKTÖRLÜĞÜ  
SOSYAL VE BEŞERİ BİLİMLER ETİK KURUL BAŞKANLIĞI  
Eğitim Bilimleri Birim Etik Kurulu  
ERZURUM

Toplantının Mahiyeti :Etik Kurul  
Toplantının Tarihi :25.05.2023  
Toplantının Sayısı :06

**Karar-18:** Danışmanlığını Doç. Dr. Merve GEÇİKLİ'nin yürüttüğü yüksek lisans öğrencisi Somayyeh SHABESTANI'nın "Makina Öğrenim Algoritmaları İle İngilizce Metin Zorluk Seviyelerinin Sınıflandırılması: Bir Yazılım Önerisi" isimli yüksek lisans tez çalışması ile ilgili etik kurul onay belgesi talebi ile ilgili husus görüşüldü

Yapılan görüşmelerden sonra adı geçen "Makina Öğrenim Algoritmaları İle İngilizce Metin Zorluk Seviyelerinin Sınıflandırılması: Bir Yazılım Önerisi" isimli tez çalışması ile ilgili çalışmaların gerçekleştirilmesinde etik ve bilimsel yönden sakınca bulunmadığına,

oy birliği ile karar verilmiştir.

(İMZA)  
Prof. Dr. Ufuk ŞİMŞEK  
Birim Etik Kurul Başkanı

(İMZA)  
Prof.Dr. İhsan Sabri BALKAYA  
Birim Etik Kurul Başkan Yardımcısı

(İMZA)  
Prof. Dr. Betül ASLAN  
Birim Etik Kurul Üyesi

(İMZA)  
Prof.Dr. Muhsine BÖREKÇİ  
Birim Etik Kurul Üyesi

(İMZA)  
Prof.Dr. Mustafa CİHAN  
Raportör

T.C.  
ATATÜRK ÜNİVERSİTESİ  
SOSYAL VE BEŞERİ BİLİMLER ETİK KURULU  
Eğitim Bilimleri Birim Etik Kurulu Karar Formu

|                          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |                                    |
|--------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------|
| <b>KARAR BİLGİLERİ</b>   | <b>Toplantı Sayısı: 06</b><br><b>Karar No: 18</b>                                                                                                                                                                                                                                                                                                                                                                                                                        | <b>Toplantı Tarihi: 25.05.2023</b> |
|                          | <p>Aşağıda bilgileri verilen yüksek lisans tez ile ilgili çalışmanın, etik ilkeler açısından değerlendirilmesi isteği ile ilgili husus görüşüldü.</p> <p>Yapılan görüşmelerden sonra; söz konusu yüksek lisans tez ile ilgili yapılacak çalışma için, araştırmanın gerekçe, amaç, yaklaşım ve yöntemleri dikkate alınarak konuyla ilgili çalışmanın gerçekleştirilmesinde etik ve bilimsel yönden sakınca bulunmadığına, Etik Kurulu oy birliği ile karar vermiştir.</p> |                                    |
| <b>ÇALIŞMA BİLGİLERİ</b> | <p><b>Proje Yürütücüsü:</b> Somayyeh SHABESTANI</p> <p><b>Çalışma Konusu:</b> Makina Öğrenim Algoritmaları İle İngilizce Metin Zorluk Seviyelerinin Sınıflandırılması: Bir Yazılım Önerisi</p>                                                                                                                                                                                                                                                                           |                                    |

| EĞİTİM BİLİMLERİ BİRİM ETİK KURULU |                              | İMZA |
|------------------------------------|------------------------------|------|
| Prof. Dr. Ufuk ŞİMŞEK              | Etik Kurul Başkanı           |      |
| Prof.Dr. İhsan Sabri BALKAYA       | Etik Kurul Başkan Yardımcısı |      |
| Prof. Dr. Betül ASLAN              | Etik Kurul Üyesi             |      |
| Prof. Dr.Muhsine BÖREKÇİ           | Etik Kurul Üyesi             |      |
| Prof. Dr. Mustafa CİHAN            | Etik Kurul Raportörü         |      |

## **CURRICULUM VITAE**

### **Personal Information**

Name - Surname: Somayyeh SHABESTANI

### **Educational Background**

M.A. Degree: Ataturk University, Institute of Educational Sciences, Department of English Language Education, 2020-2023

Bachelor's Degree: Islamic Azad University, Institute of Educational Sciences, Department of English Language Education, Maragheh-Iran, 2009-2011

### **Work Experience**

Since I won first place in the university entrance exam in 2009, I taught English in different language schools for all ages simultaneously.

I served as the head teacher at Erzurum American Culture School for three years between 2017 and 2020. My teaching career was very rewarding for me in terms of experience.