,

Birzeit University

Faculty of Engineering and Technology

Department of Electrical and Computer Engineering

Graduation Project ENCS5300

Automated Writing Evaluation in Arabic Language

Prepared by:

Diaa Maali

Aseel Khaseeb

Zainab Shaabneh

Supervised by: Dr. Abualseoud Hanani

Section:4

A graduation project submitted to the Department of Electrical and Computer Engineering in partial fulfillment of the requirements for the degree of B.Sc. in Computer Engineering.

Abstract

Although the manual evaluation of essays is a time-consuming process, writing essays has a significant role in assessing learning outcomes. Therefore, automated essay evaluation represents a solution, especially for schools, universities, and testing companies. Moreover, the existence of such systems overcomes some factors that influence manual evaluation such as the evaluator's mental state, the disparity between evaluators, and others. In this paper, we propose an Arabic essay evaluation system based on Gradient Boosting Classifier along with a wide range of features including morphological, syntactic, and semantic features. The system evaluates essays according to five criteria: spelling, coherence level, and style, without the need for domain-representative essays (a model essay). A specific model is developed for each criterion; thus, the overall evaluation of the essay is a combination of the previous criteria results. We develop our dataset based on essays written by our university students whose native language is Arabic. then The dataset is evaluated by experts. The experimental results and the correlation between the system and the experts' evaluation is 70%. Additionally, the system shows variant results in evaluating criteria separately. A representative set of 20 features was used across all of the reported experiments for fairness and comparability. five different classifiers were trained and then used to assess the testing writing texts. The results showed that the Gradient Boosting Classifier achieved the highest accuracy (0.70) due to its boosting technique. The model's accuracy (0.71) surpassed some systems and approached that of the multiple linear regression method (0.77). Overall, the research successfully explored various models and highlighted the importance of complexity, parameter tuning, and data handling for accurate classification.

المستخلص

على الرغم من أن التقييم اليدوي للمقالات هو عملية تستغرق وقتًا طويلاً ، إلا أن كتابة المقالات لها دور مهم في تقييم نتائج التعلم. لذلك ، يمثل التقييم الآلي للمقال حلاً ، خاصة للمدارس والجامعات وشركات الاختبار. علاوة على ذلك ، فإن وجود مثل هذه الأنظمة يتغلب على بعض العوامل التي تؤثر على التقييم اليدوي مثل الحالة العقلية للمقيم ، والتفاوت بين المقيمين ، وغيرها. في هذا البحث ، نقترح نظامًا لتقييم المقالات العربية يعتمد على تصنيف تعزيز التدرج جنبًا إلى جنب مع مجموعة واسعة من الميزات بما في ذلك السمات المورفولوجية والنحوية والدلالية. يقوم النظام بتقييم المقالات وفقًا لخمسة معايير: التهجئة ومستوى التماسك والأسلوب ، دون الحاجة إلى مقالات تمثيلية عن المجال (مقال نموذجي). يتم تطوير نموذج محدد لكل معيار ؛ وبالتالي ، فإن التقييم العام للمقال هو مزيج من نتائج المعايير السابقة. نقوم بتطوير مجموعة البيانات الخاصة بنا بناءً على المقالات التي كتبها طلاب جامعتنا لغتهم الأم هي العربية. ثم يتم تقييم مجموعة البيانات من قبل خبراء. النتائج التجريبية والعلاقة المتبادلة بين النظام وتقييم الخبراء ٧٠٪. بالإضافة إلى ذلك ، يُظهر النظام نتائج متباينة في تقييم المعايير بشكل منفصل. تم استخدام مجموعة تمثيلية من ٢٠ ميزة في جميع التجارب المبلغ عنها لتحقيق الإنصاف وقابلية المقارنة. تم تدريب خمسة مصنفات مختلفة ثم استخدامها لتقييم اختبار كتابة النصوص. أظهرت النتائج أن مصنف تعزيز التدرج حقق أعلى دقة (٧٠.٠) بفضل تقنية التعزيز الخاصة به. تفوقت دقة النموذج (٧١.٠) على بعض الأنظمة واقتربت من دقة طريقة الانحدار الخطي المتعدد (٧٧.٠). بشكل عام ، نجح البحث في استكشاف نماذج مختلفة وسلط الضوء على أهمية التعقيد وضبط المعلمات ومعالجة البيانات من أجل التصنيف الدقيق.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation and Overview

The writing skills evaluation is one of the important processes in the education system, especially in language learning. In most education stages, it is necessary to develop writing and presentation skills, such as language proficiency and creativity. However, the traditional way of writing evaluation is done manually, which requires a great effort and a long time, especially when there is a large number of essays/articles to be evaluated. This process has been made easier with the progress in Natural Language Processing (NLP) technology, or the so-called Automated Essay Evaluation systems (AES). As the automated writing skills evaluation also overcomes some of the factors related to the manual method such as the mental state of the rater, biases, and disparity among the raters. Moreover, automated essay evaluation systems are tools that can help new teachers and learners to train and improve their writing skills. It also helps in reducing the heavy burden of assessment on teachers.

In this project, we are trying to use the latest technologies in natural language processing and deep learning to develop an automatic intelligent system that can do the writing skills evaluation in the Arabic language. Writing skill is usually measured by various methods, such as asking learners (e.g. students) to write an essay or article about a specific subject, or asking learners to present an answer to a specific question; an open question, or a topic-specific question.

Numerous studies on AES have been conducted for the English language, and commercial programs have been developed for use in English learning institutions. In contrast, it appears that Arabic AES systems are restricted to short-answer questions with pre-selected teacher-selected response formats. This constraint exists due to the complexity of the Arabic language and the dearth of Arabic (NLP) tools. To fill this gap, we present a grading system for Arabic essays in this study that uses machine learning without the need for an essay model.

1.2 Applications

AES systems can be applied in schools and universities to help teachers in evaluating their students' writing skills. It also can be applied in the language scale exams in institutions and the employment process. It can help editors, journalist publishers, and essay writers by addressing some of the issues with traditional evaluation.

1.3 Problem Statement

The traditional way of writing g skill evaluation is done manually by an expert/teacher who needs to read the writing presentation submitted by the learner/student carefully and use evaluation criteria to evaluate each writing and give a useful data set. This way has some limitations, such as it requires a great effort of concentration and consumes a lot of time. In addition, this method is a subjective method, which means that it is highly affected by the thoughts and the mode of the evaluator. For example, the same essay can be evaluated as good by one evaluator and bad by another, or even different assessments by the same evaluator in different environments. Presenting a system that can do this evaluation automatically will overcome most of these limitations and make it easier for the learners to practice writing and improve their writing skills by getting useful feedback. Moreover, automating the writing skills evaluation makes it easier to make online exams with essay questions, where students are asked to present their answers in a free text without any constraints.

1.4 Project Objectives

The main objectives of this project are listed below:

- review sufficient related and recent studies in the field of writing skills assessment.
- Collect sufficient samples of written essays by students in the Arabic language.
- Get the assessment of each essay done by a language expert, usually, by the student's teacher.
- Analyze the collected data set and prepare it for developing our proposed system, which can do the assessment automatically.
- Use state-of-the-art techniques in natural language processing and machine learning techniques for building the proposed system.
- Conduct a various number of experiments using the collected data set and the proposed methodology.
- Compare the proposed system performance with other similar systems as reported in the published studies.

2 Related Work

AES is a worthwhile endeavor that can aid in the advancement of automated assessment and assist teachers in lightening the heavy workload associated with assessment. More and more scholars are starting to focus on his area as a result of the growth of online education in recent years. This section briefly highlights several studies that he been conducted to attempt to establish automated evaluation systems. While reviewing the related studies, we found that most of the data sets had studies focused on the English language, and very few studies have been published in the Arabic language. Therefore, this section is divided into two subsections; related studies revision in Arabic and English languages.

2.1 Automatic writing assessment in the Arabic language

In 2014, Alghamdi et. al. in [1], applied the linear regression technique to forecast anyone-paragraph. They used Latent Semantic Analysis (LSA), the number of words, and spelling errors. The 579 essays that were gathered from undergraduate university students and graded by two professors were subjected to the suggested system. According to the results, 96. cites percent of the essays were accurately evaluated, and the correlation between automated and manual grading was 0.78, which is close to the value of 0.7 attained by inter-human correlation.

In 2019, Thamer Al-Rousan et. al. in [2] suggested that at the present stage, automated essay scoring (AES) cannot replace human raters in essay grading tasks but can serve as a valuable tool for assisting human raters as secondary evaluators. Most of the existing AES systems have been developed in Western countries, and there is currently no commercial AES system available in the Asian region. The literature indicates that the majority of studies have focused on assessing essays based on their content using the Latent Semantic Analysis (LSA) technique. Additionally, many reported implementations treat AES as a supervised document classification task. Building upon this literature review,[2] proposed three types of supervised general frameworks: content similarity, machine learning, and hybrid approaches. They also introduced a new framework that evaluates essays based on both content and linguistic features It is worth noting that different AES research studies have employed various evaluation methods to assess the performance of proposed models, making it challenging to conduct proper com-

parative studies. To address this issue, the Quadratic Weighted Kappa in [2] as a standardized method for evaluating AES performance. This proposed method aims to promote consistency and uniformity in the development of AES systems.

In 2020, Abeer Alqahtani et. al. in [3] proposed a support vector regression (SVR) algorithm-based method for evaluating Arabic essays that uses features from various linguist levels. Independent tests were done to predict the five criteria scores: punctuation, spelling, structure, and coherence. Combining the results of the preceding criteria scores, a holistic score for the essay was determined. The 200 essays that make up the used dataset. The proposed system performed 96 percent accurately overall and had a 0.87 correlation with manual evaluation; however, it only performed 77 percent accurately for spelling, 91 percent for structure, 87 percent accurately for coherence, 78 percent accurately for style, and 93 percent for punctuation.

In 2021, Wee Sian Wong et. al. in [4] focused on improving the accuracy of the automated article system for human matching results. The generated dataset contains 120 questions with 3 sample answers for every question. In addition, the dataset used in this study was provided by Kaggle datasets [1]. The results of the automated grading of the essays showed which is a combination of Arabic Word-Net (AWN) produces a better result compared to the case without Use Word-Net based on the mean absolute error value and Pearson's correlation. As a result of this research, there is an outlook for future work on the use of the machine and neural learning grid models to enhance the accuracy of the classification of Arabic articles, in addition to studying the impact of Word embedding technology.

In 2022, Ramesh Dadi et. al. in [5] empathize that AES is an essential educational app Using natural language processing and deep learning. Although current AES systems fail in certain areas Like content-based assessment, paradigms are not tested on hostile responses and make comments on the article. Although current AES systems fail in certain areas Like content-based assessment, paradigms are not tested on hostile responses and make comments on the essay. In this study, the proposal is made on a sentence basis embedding to capture text cohesion and coherence in a vector. And they worked on these vectors on Long Short-Term Memory (LSTM) and Bidirectional LSTM to find the sentence link internally and externally

[1] https://www.kaggle.com/general/260690

to check sentence framing and its association with other sentences, and its importance to the mentor.

2.2 Automatic writing assessment in English

In 2003, Jill Burstein et. al in [6]The suggested writing analysis tool aims to detect grammar errors by categorizing them into five main types: agreement errors, verb formation errors, incorrect word usage, missing punctuation, and typographical errors. To achieve this, two approaches are utilized: a corpus-based approach and a statistical-based approach. The tool is trained on a large collection of edited text, from which it extracts and tallies sequences of adjacent word and part-of-speech pairs called bi-grams. By comparing the occurrence of these bi-grams in student essays with the expected frequencies based on the corpus, the tool identifies and flags instances where they deviate significantly. Determining what constitutes good or bad writing style is subjective since individuals have different preferences. Nonetheless, the proposed tool highlights certain style aspects that the writer might want to revise. This includes the use of passive sentences, excessively long or short sentences, and repetitive words. These elements can impact the overall quality assessment of the essay.A well-crafted essay should incorporate various components of discourse, such as introductory material, a problem statement, main ideas, supporting ideas, and a conclusion. The system's performance is evaluated based on precision and recall metrics.

In 2019 Nicky Hockly in [7] introduced a set of techniques outlined in Griffith's work (Hockly, 2012) for the automated analysis of writing. The proposed methodology utilizes a combination of statistics, natural language processing (NLP), artificial intelligence (AI), and machine learning to evaluate constructed written responses across various aspects including grammar, syntactic complexity, mechanics, style, topical content, content development, and deviance.

The approach primarily focuses on examining the surface-level linguistic features of the text. It is based on the concept of trains and proxies, where trains represent intrinsic variables such as grammar (parts of speech, sentence structure), fluency (essay length), and diction (variation in word length). By analyzing these features, the methodology aims to provide a comprehensive assessment of the written text.

In 2020 [8] With the Internet's rapid expansion, more and more people are using social media platforms to express their thoughts and viewpoints. Among these platforms, Twitter has gained immense popularity, attracting millions of active users who share information through tweets. To analyze the sentiments and opinions conveyed in these tweets, the field of sentiment analysis (SA) has emerged.A novel approach has been proposed for SA on Twitter data, utilizing a gradient boosted decision tree (GBDT) classifier. This method surpasses existing deep learning techniques in terms of performance. The results highlight its effectiveness in analyzing sentiment within large-scale data sets. As for future advancements, potential extensions could involve experimenting with a comprehensive thesaurus and assessing the approach's applicability to diverse languages.

In 2022 Ramesh Dadi in [9]A substantial dataset consisting of over 1500 essays was employed in this study, with human raters ranging from 20 to 60. To evaluate the essays, an automated rater was utilized, employing a single-dimensional LSTM (Long Short-Term Memory) model. The assessment process involved K-fold cross-validation on the ASAP dataset, which can be found at the provided link (https://www.kaggle.com/c/asap-aes).

In order to prepare the word vector, the Word2Vec NLP library was utilized. This article presents an improvement in model accuracy and incorporates word2vec to extract features from the articles. However, it's worth noting that these features are extracted at the word level, which introduces the possibility of losing the article's semantics. Maintaining the article's semantics is a crucial aspect in article classification, and there is a potential risk of it being compromised in our model.

3 Methodology

3.1 Data set collection

Our data set consists of written essays submitted as answers in the online exam by students in the Arabic language course at Birzeit University during the period 2020/2021 when the COVID-19 pandemic happened and the teaching was turned to online. The essays were about three different topics/questions. This made our work easier because we did not have to re-enter the handwritten texts into the computer as in the traditional handwritten exams. Instead, we get the text in CSV format, as shown in figure 3.1 which is ready to be used in the e-learning system. Our data set contains 570 essays with an average length of about 250 words (3KB) on the three topics. The questions were answered by 570 students, all students answered the three questions. The first two questions have the same weight of 8 marks and the third question has a different weight of 14 marks, so, the overall grade is 30. About 14 students did not answer any questions, they were graded zero. Five students answered only one question, leaving two fields unanswered; so we have 1658 answers for the three topics together.

I	J	K
Response 1	Response 2	Response 3
الموضوع الأول: التعليم الالكتروني، والتعليم الوجاهي، ال	يوجد الكثير من الغابات المنتشرة في مختلف قارات العالم	كورونا والأزمة الاقتصادية وأخذ الحذر
ظاهرة المركبات غير المرخصة ادت إلى تذبذب القيم والأ	ه في الواقع، سرعان ما يتلاشى ويتوارى عن الأنظار "	يعد فيروس كورونا من أخطر الفيروسات التي أثرت بالعالم
في وقتنا الحاضر، بعد ظهور مرض الكورونا ، أصبح	أستطيع أنْ أتخيل في عقلي صورًا مختلفة صغيرة وكبيرة،	يعد فيروس كورونا كوفيد 19 من أشدّ الفيروسات التاجيّة
إن أفضل الطرق التي توصل اليها الانسان للوصول إلى الت	الغابة السوداء أشهر الغابات في ألمانيا وأجملها.	مر العالم بالكثير من الكوارث التي أثقلت كاهله. ابتداءً من
الطاقة الشمسية وطاقة الرّياح من أهم أنواع الطاقة المتج	ذي لا وجود له في الواقع، سرعان ما يختفي من الذاكرة "	يعد فيروس كورونا كوفيد19 من أكثر الامراض فتكاً وتاثي
في الآونة الأخيرة، ومع ظهور وباء كورونا، أصبح الزاماً عل	يوجد الكثير من الغابات المنتشرة في أرجاء العالم، والكثير	تطورت العصور والاجيال، وتطورت معها الابتكارات والاخ
ادى انتشار ظاهرة المركبات غير المرخصة في الريف الف	استطيع ان اتخيل في عقلي وذهني صورا مختلفة ومتعدد	تعرض العالم لضربة قوية شلت قطاعاته بكافة اشكالها بع

Figure 3.1: Some samples from the Dataset

3.2 Data set evaluation

The course instructor, Manal Hassan who has a Ph.D. in Arabic language and works at Birzeit University, evaluated the collected student's answers. The evaluation was made according to a specified criteria and evaluation model which include: spelling mistakes, grammar errors, structure, coherence, level of coherence, style, and punctuation. The instructor was concerned with the proper spelling of words. She identified each spelling error and categorized it into one of four categories: errors on hamza/, replacement letters, additional letters, or omitting letters. she looked for the presence of four key components for the structural criterion: the title, the introduction, the body, and the conclusion. She assessed the coherence between the essay's parts and the title and the cohesion within each section. This criterion also has to do with using connectives correctly. She took into account word repetition, sentence length, word choice, and avoiding long speeches when evaluating an essay's style. She also assessed punctuation marks by Arabic norms. Taking all of these into account, the total exam score was 30 marks. The distribution of students' marks was as shown in the following table 3.1.

Grade	Category	Percentage of students who got this grade
(25-30)	A	30%
(20-24)	B	55%
(10-19)	C	10%
(less than 10)	D	5%

Table 3.1: Distribution of student marks

To increase confidence in the manual evaluation, we re-evaluated the entire data again with a group of Arabic language experts and got another evaluation for each text. To evaluate the level of agreement between evaluators Cohen's kappa coefficient was calculated. Cohen's Kappa coefficient is a widely used method to evaluate inter-rater agreement levels for categorical scales, and it calculates the proportion of agreement that is chance-corrected. The online kappa calculator used [2]. The test achieved a level of agreement measured between the evaluators +0.893, which is considered almost perfect agreement according to Fleiss [10] values larger than +0.75 are characterized as excellent agreement.

The details of the Kappa test are shown in figure 3.2.

	A	B	C	D	Total
A	100	12	0	0	112
B	6	300	5	0	311
C	0	4	66	6	76
D	0	0	3	50	53
Total	106	316	74	56	552

Number of observed agreements: 516 (93.48% of the observations)
Number of agreements expected by chance: 215.1 (38.97% of the observations)

Kappa= 0.893
SE of kappa = 0.017
95% confidence interval: From 0.860 to 0.927

Figure 3.2: Kappa-Test

[2] https://www.graphpad.com/quickcalcs/kappa1/

3.3 Language writing assessment criteria

Our proposed system depends on different representative features retrieved by considering the standards people use when evaluating writing skills. We worked on a wide range of features represented by numbers at various levels by the system output, which is numeric scores, including:

A. Surface features: Only text-related features, including word frequencies, are shown in the surface features. An excellent essay does not use a lot of synonyms or informal language, nor does it repeat words. Instead, it ought to use effective word choice and a variety of sentence lengths.

B. Structure features: The title, introduction, body, and conclusion are the four sections of a well-organized essay.

C. Coherence features: The coherence criterion is used to assess how closely essay parts link to the title, how cohesively essay parts fit together, how well acceptable discourse connectives are used, and how diverse the connectives are.

D. Spelling features: One of the most crucial criteria used while correcting is spelling. This is an example of a probable clerical error that alters text meanings.

E. Punctuation marks: Punctuation marks indicate proper, improper, and missing word usage as well as discourse connectives. They divide the essay into clauses and sentences. By doing so, you might be able to see some errors, including missing commas. Possible errors refer to punctuation that is not followed by conjunction or discourse connective since each punctuation mark serves a specific purpose (e.g., a full stop used after a phrase or a comma used within a sentence to break it into clauses). Errors on hamza, replacement letters, additional letters, and omission letters are the four categories into which they fall. Also, we concentrated on using verbs, prepositions, and adverbs correctly.

3.4 System Overview

In this section, we provide our methodology to achieve the objectives of this project. First, the essays need to be tokenized into sentences, and then each sentence is tokenized into words. Some special symbols like @, #, etc. are removed from the essays. In the next stage, each criterion in the evaluation metric (spelling, structure, coherence, etc) needs to be extracted from the text, at the sentence level and word level. There are various techniques and algorithms to extract the representative features for each criterion. For example, the state-of-the-art word embeddings pre-trained models, which are based on deep learning technology, will be used to extract some of these features. More specifically, in Universal Sentence Encoder (USE) [11], each sentence is represented by a vector, which represents the semantic features of that sentence. Using USE, each phrase in an essay is broken down into its separate words, bi-gram, and then embedded as a vector. The vectorized words are then averaged to produce a sentence-level vector, which is then transmitted to feed-forward deep neural networks for sentence embedding, as shown in the figure 3.3

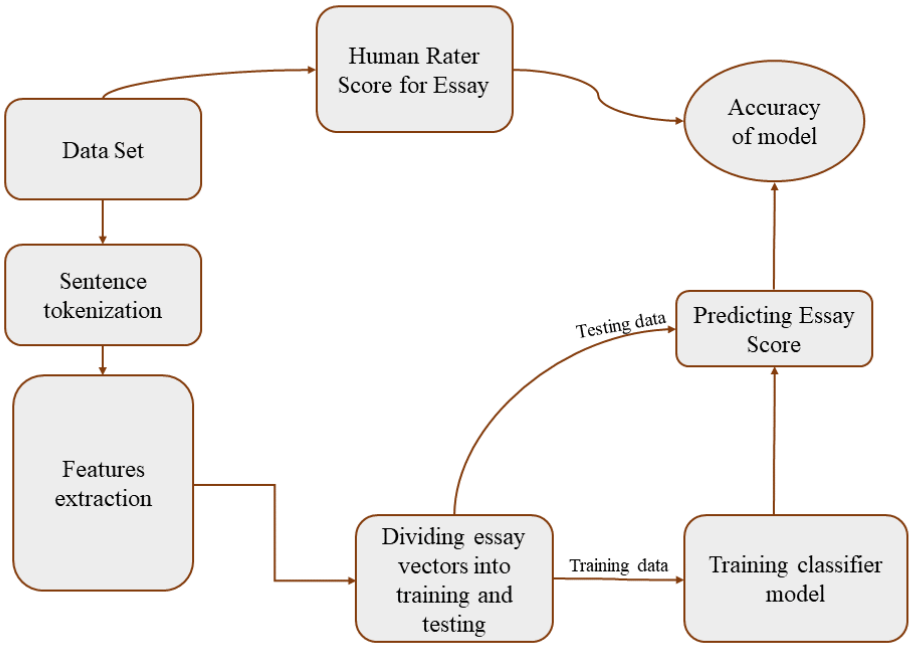

Figure 3.3: Overview of proposed AES system

3.4.1 Tokenization

Tokenization is an important preprocessing step that can help improve any NLP-based model's performance. It is a step based on dividing the text into sentences or words, where special symbols used a sentence boundary such as (", ".", "?" or "!") can assist this process. In Arabic, tokenization can be a challenge due to the complexities of the language, such as the presence of diacritics and the use of the non-Latin script[12].

3.4.2 Features extraction

The main aim of feature extraction is to represent each criterion of the writing assessment with a feature, or a set of features that can be used in machine learning modeling. In the context of writing assessment, feature extraction is a process that involves identifying the topic sentence and supporting details in each paragraph, analyzing the types of transitions used, examining the choice of vocabulary, and analyzing the use of grammar and sentence structure. Additionally, the conclusion of the essay is also an important feature to consider. By extracting these features, one can gain insight into the organization, coherence, and style of the essay, we try a wide range of features represented by numbers at different levels including:

3.4.3 Syntactic and morphological features

Syntactic and morphological features are two types of linguistic features commonly used in natural language processing (NLP) to analyze text. Syntactic features refer to the arrangement of words in a sentence, including the structure and function of phrases and clauses. Morphological features, on the other hand, refer to the study of the internal structure of words, including the way they are formed and the meaning conveyed by their various parts. Some examples of syntactic features include part-of-speech (POS) tags, dependency parse trees, and constituency parse trees. POS tags identify the grammatical category of a word, such as noun, verb, or adjective, while dependency and constituency parse trees represent the syntactic structure of a sentence. Morphological features, on the other hand, include information about word inflections, such as tense, number, and gender. These features can be used to disambiguate the meaning of a word, particularly in languages with rich inflectional systems. Syntactic and morphological features are particularly useful in NLP tasks such as text classification, information extraction, and machine translation. For example, POS tags can be used to identify the subject

and object of a sentence, which can be useful for tasks such as sentiment analysis or named entity recognition. Morphological features can be used to disambiguate the meaning of a word, which can improve the accuracy of tasks such as text classification or machine translation.

3.4.4 Surface features

Surface features refer to observable characteristics of text that are typically extracted from the surface of the language, without a deeper analysis of its meaning. These features include various aspects of the text, such as its vocabulary, syntax, and morphology. Some examples of surface features include word frequency, word length, sentence length, part-of-speech tags, and punctuation marks. Surface features can be useful in a variety of NLP tasks, such as text classification, authorship attribution, and sentiment analysis. For example, word frequency can be used to identify the most common words in a document, which can be used to determine the topic or genre of the text. Sentence length can be used to identify complex or difficult-to-read passages in a text, which may indicate the need for simplification or clarification.

3.4.5 Discourse features

Discourse is a type of linguistic feature used in natural language processing (NLP) to analyze text beyond the level of the individual sentence. Discourse features refer to the relationships between sentences and how they contribute to the overall meaning of a text. Some examples of discourse features include rhetorical structure, coherence, cohesion, and discourse markers. Rhetorical structure refers to how a text is organized, such as the use of headings or subheadings, the use of transitions between sections, and the overall flow of the text. Coherence refers to the extent to which a text is internally consistent and logically connected, while cohesion refers to the linguistic devices used to connect sentences and paragraphs, such as pronouns, conjunctions, and lexical repetition. Discourse markers, on the other hand, are words or phrases used to signal relationships between sentences, such as contrast, cause and effect, or concession. Some examples of discourse markers include "however", "therefore", "because", and "although". Discourse features are particularly important in NLP tasks such as text summarization, document classification, and dialogue analysis.

3.4.6 Spelling features

The FARASA spell checker [13] classifies the mistakes that FARASA provided into four classes: mistakes on 'hmza', replacement letters, extra letters, and omission letters. Because the model is built almost entirely on FARASA features, We examined our dataset with this tool to determine how frequently FARASA detects actual spelling errors. The number of cases in which FARASA identifies the correct word as incorrect, which were 2130, 85, and 250 incidents, respectively. However, FARASA only recovered 120 words because, in some circumstances, it can be difficult for humans to notice the missing spaces.

3.4.7 Coherence features

The degree to which essay parts are related to the title, the cohesion between essay parts, the usage of the proper discourse connectives, and the variety of connectives are all evaluated using the coherence criterion. Consequently, we included lexical features, syntactic features, and surface features. The majority of these qualities are generic and contain details on the many components of essays (lengths, basic syntactic traits, etc.). These characteristics are used to determine essay structure, which helps to determine the degree of acceptable coherence.

3.4.8 Style features

A well-written essay avoids using the same words over and over again without substituting synonyms avoids using colloquial language, and makes appropriate use of a variety of sentence lengths. We used surface features, which primarily look at sentence and paragraph length, lexical features, which look at punctuation usage, and word count, which may have an impact on style. To further explore synonymy, we also include semantic traits. However, the number of discourse connectives and the number of synonyms throughout the entire essay.

3.4.9 Sentiment analysis

is an NLP technique used to determine the emotional tone of a paragraph? It involves analyzing text to classify it as positive, negative, or neutral. The process includes preprocessing the text, extracting relevant features, training a classification model, and predicting the sentiment of new paragraphs. Sentiment analysis is widely applied in areas like social media monitoring, customer feedback analysis, and market research. While it's not always perfect due

to language nuances, it provides valuable insights into the subjective information conveyed in the text, aiding decision-making processes.

3.5 Classification models

3.5.1 ANN

An Artificial Neural Network (ANN) classifier is a type of machine-learning model inspired by the structure and functioning of the human brain. It is designed to recognize patterns and make predictions based on input data. The network consists of interconnected nodes, known as neurons, organized in layers (input, hidden, and output). Each neuron processes input data using weights and biases, and the network learns to adjust these parameters during training to improve its accuracy.

The training process involves feeding the network with labeled data, and it adjusts its internal parameters to minimize the difference between predicted outputs and the true labels. Once trained, the neural network can be used to classify new, unseen data into predefined categories, making it a powerful tool for various tasks like image recognition, natural language processing, and sentiment analysis. Neural network classifiers are popular due to their ability to handle complex and high-dimensional data, but they require substantial computational resources and data for effective training.

3.5.2 Decision Trees

It is a popular supervised machine learning algorithm that can be used for both classification and regression tasks. They work by recursively splitting the input data based on different feature attributes to create a tree-like structure of decision rules. Each internal node in the tree represents a decision based on a specific feature, while each leaf node represents a class label or a predicted value. The decision-making process follows a hierarchical structure, where the input data traverses down the tree from the root node to a leaf node, resulting in a predicted outcome. Decision Trees are advantageous as they are easy to understand and interpret, can handle both categorical and numerical data, and can capture complex decision boundaries. However, they can be prone to overfitting if not properly controlled, and they may struggle with handling high-dimensional or noisy datasets.

3.5.3 Random Forest

is an ensemble learning method that combines multiple decision trees to make predictions. It is a versatile and powerful algorithm used for both classification and regression tasks. The key idea behind Random Forest is to create a collection of decision trees, each trained on a different subset of the training data and using a random subset of features. During prediction, each tree in the forest independently generates its own prediction, and the final prediction is determined by aggregating the results from all the trees. This aggregation can be done by taking a majority vote in the case of classification or averaging the predictions for regression. Random Forest is highly effective in handling high-dimensional data, dealing with missing values, and mitigating overfitting. It is known for its robustness, scalability, and ability to provide insights into feature importance.

3.5.4 SVM

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm used for binary classification and regression tasks. It finds an optimal hyperplane that best separates data points of different classes in a feature space, maximizing the margin between them. SVM can handle non-linearly separable data by mapping it to a higher-dimensional space using kernel functions. The algorithm is effective in high-dimensional spaces and is widely used in various applications. While SVM generalizes well and can handle complex data, it may be computationally expensive and requires careful tuning of hyperparameters and kernel selection.

3.5.5 Gradient Boosting Classifier

It is a powerful machine-learning algorithm that belongs to the ensemble learning family. It is specifically designed for classification tasks and has gained significant popularity due to its effectiveness in handling complex datasets. The algorithm works by combining multiple weak classifiers, typically decision trees, into a strong predictive model. Each weak classifier is trained sequentially to correct the mistakes made by the previous models, thereby improving the overall prediction accuracy. By iteratively optimizing a loss function through gradient descent, the Gradient Boosting Classifier gradually learns to make better predictions. It excels in capturing intricate relationships between features, handling noisy data, and preventing overfitting. With its ability to handle diverse data types and high-dimensional feature spaces, the Gradient

Boosting Classifier has become a go-to choice for a wide range of classification tasks in various domains.

3.6 System performance metric

In all of the conducted experiments presented in this report, four performance measures were calculated to represent the system performance, which are:

1)Accuracy: It measures the overall correctness of a classification model. It calculates the ratio of correctly predicted instances to the total number of instances in the dataset.
Mathematical Equation:

$$Accuracy = \frac{TruePositives + TrueNegatives}{TruePositives + TrueNegatives + FalsePositives + FalseNegatives} \quad (1)$$

Importance:
Accuracy provides a general understanding of how well a model performs in terms of correctly predicting both positive and negative instances. However, it might not be suitable for imbalanced datasets, where one class is significantly more prevalent than the other.

2)Recall(Sensitivity or True Positive Rate): Recall measures the ability of a model to identify all positive instances correctly. It calculates the ratio of true positives to the sum of true positives and false negatives.
Mathematical Equation:

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \quad (2)$$

Importance:
The recall is important in scenarios where identifying positive instances is crucial, such as identifying disease cases or detecting fraudulent transactions. A high recall indicates that the model can effectively minimize false negatives, reducing the chances of missing important positive instances.

3)Precision(Positive Predictive Value): Precision measures the ability of a model to correctly

identify positive instances out of the total predicted positive instances. It calculates the ratio of true positives to the sum of true positives and false positives.
Mathematical Equation:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{3}$$

Importance:
Precision is important when the cost of false positives is high. It ensures that the positive predictions made by the model are highly likely to be correct. For example, in email spam detection, precision is valuable as it reduces the chances of legitimate emails being classified as spam.

4)F1-score: The F1-score combines both precision and recall into a single metric. It provides a balanced measure by calculating the harmonic mean of precision and recall.
Mathematical Equation:

$$F1score = \frac{2 * ((Precision * Recall)}{Precision + Recall} \tag{4}$$

Importance:
The F1-score is particularly useful when there is an uneven class distribution or when both precision and recall need to be considered together. It provides a single value that summarizes the performance of a classification model, taking into account both false positives and false negatives.

These metrics help evaluate the performance of machine learning models, allowing practitioners to assess their effectiveness in various scenarios.

4 Experiments and Results

4.0.1 Experimental setup

In all presented experiments, the collected data was split into two sub-sets; 90% for training and 10% for testing. Since the dataset is relatively small, the 10% cross-validation technique was used by repeating the same experiment by replacing the selected 10% testing essays with new ones and adding the current to the training data. A set of represented features were extracted from each essay after applying word, sentence, and paragraph tokenization. Some features are computed at the word level, some at the sentence level, and some are computed at the paragraph level.

Various machine learning models were trained on the features extracted from the training dataset and then used to do the essay evaluation by classifying the given essay from the testing subset into one of the evaluation categories; A, B, C, or D. All experiments were implemented in Python using the Collab online workspace [3], offered by Google. The following are the most important libraries that we used in the presented experiments:

- from spellchecker import SpellChecker
- from nltk.tokenize import word$_tokenizefromfarasa.segmenterimportFarasaSegmenter$
- from farasa.pos import FarasaPOSTagger
- from nltk.tag import pos$_tagfromnltk.tokenizeimportword_tokenize, sent_tokenize$
- from sklearn.neural$_networkimportMLPClassifierfromsklearn.ensembleimportGradientBoosting$
- from sklearn.svm import SVC
- from sklearn.ensemble import RandomForestClassifier
- from sklearn.metrics import accuracy$_score, recall_score, precision_score, f1_scoreimportlanguage_tool_py$
- from gensim.models import CoherenceMode

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

4.0.2 Features computation

numWordsList: Number of words in each text.

** numWordsList = len(words)

numSentencesList: Number of sentences in each text.

** **Equation**: numSentences = len(sentences)

readabilityList: Readability score for each text, calculated using a formula involving the number of words and syllables.

** readability = 0.39 * (num-words / num-sentences) + 11.8 * (syllables / num-words) - 15.59

sentimentList: Sentiment score for each text using the SentimentIntensityAnalyzer from NLTK.

** The sentiment score is obtained using the SentimentIntensityAnalyzer, and the specific calculation is not shown in the provided code. The score is typically based on a pre-trained model that considers the positive, negative, neutral, and compound (overall) sentiment of the text.

avgwordLengthlList: Average word length for each text.

$avgWordLength = \frac{totalLength}{numWords}$

where total length is the sum of the lengths of all words in the text.

avgSentLengthList: Average sentence length for each text.

$avgSentLength = \frac{totalSentLength}{numSentences}$

where totalsentlength is the sum of the lengths of all sentences in the text.

errorsList: Number of grammar and spelling errors in each text, checked using the **language-tool-python library.** and spellchecker libraries.

errors = len(tool.check(text))

where the tool is the LanguageTool instance from the **language-tool-python library.**

type -token-ratio -list: Type-Token Ratio for each text, which is the ratio of unique words to the total number of words.

$type - token - ratio = \frac{len(unique - words)}{num - words}$

where unique words are the set of unique words in the text.

coherence-scores-list: Coherence score for each text. It seems to be calculated by measuring the similarity between each text and a predefined title using TF-IDF

vectorization and cosine similarity.

The coherence score is obtained using cosine similarity between the TF-IDF vector representation of each text and the title.

num-spelling-errors-list: Number of spelling errors in each text.

num-spelling-errors = len(misspelled)

4.0.3 Results and discussion

A consistent set of 20 features were used in all of the reported experiments, ensuring fairness and comparability among different models. The selection of these features relies on prior knowledge and existing research. By using the same feature set, a more accurate comparison between models can be achieved.

Model Name	Accuracy	Recall	Precision	F-Score
Support Vector Machine	0.52	0.62	0.66	0.62
Decision Tree Classifier	0.52	0.60	0.60	0.60
Neural Networks Classifier	0.54	0.73	0.81	0.77
Random Forest Classifier	0.66	0.64	0.66	0.64
Gradient Boosting Classifier	0.70	0.72	0.80	0.76

Table 4.1: model performances.

Five models were compared for classification, and the reasons for the lower accuracy in some models compared to others can be explained. The Random Forest Classifier showed moderate accuracy (0.66). the lower accuracy in this model could be due to its complexity and the need for appropriate settings. On the other hand, the Gradient Boosting Classifier achieved the highest accuracy (0.71) among the compared models. Boosting combines the predictions of multiple weak base models to create a strong and accurate model, leading to improved performance. As for the Neural Networks Classifier, it exhibited lower accuracy (0.54). The reason for the lower accuracy here could be the need for proper parameter tuning of the neural network and appropriate data organization to achieve better performance. The Support Vector Machine model also had a low accuracy (0.52). The reason behind this could be the difficulty in determining the optimal margin or boundaries between different classes, which affects the final accuracy of the model. Finally, the Decision Tree Classifier also showed low accuracy (0.52). The performance of this model relies on complex sequential decisions, and the reason for the

lower accuracy could be the model's ability to handle data containing significant overlaps or complexities. In summary, the lower accuracy in some models can be attributed to factors such as model complexity, inappropriate parameter tuning, difficulty in determining the optimal margin, and insufficient handling of data overlaps or complexities. In contrast, the Gradient Boosting Classifier benefits from the boosting technique to improve performance and increase accuracy. Given the results and considering the overall project context, there are compelling reasons to consider the accuracy of the Gradient Boosting Classifier as good. The higher accuracy value (0.71) indicates that the model can evaluate text essays with good accuracy, along with a decent recall (0.73). The boosting technique used in the model contributes to improving performance and increasing accuracy. Furthermore, the model's performance can be further enhanced through additional parameter tuning and addressing potential issues.

System	Approach	Result
Tirthankar Dasgupta et al.[11]	CNN , Bidirectional, LSTMs ,neural network	0.79
Wang et al. [12]	Bi-LSTM	0.73
Jiawei Liu et al. (2019)	CNN,LSTM , BERT	0.71
Darwish et al.[15]	Multiple Linear Regression	0.77
Uto(B) and Okano [17]	Item Response Theory Models (CNN,LSTM, BERT)	0.77
Zhu and Sun [16]	RNN(LSTM,Bi-LSTM)	0.71
Our best system	**ANN classifier**	**0.71**

Table 4.2: Comparison of results.

The accuracy of our model (0.71) was compared with the results reported in the reference study. With this comparison, our accuracy can be considered good compared to some other results. It outperformed the system that achieved an accuracy of 0.70 and came close to the accuracy achieved by the multiple linear regression method (0.77). We took into account the general context of the research and the specific requirements when evaluating the quality of accuracy. We consider the accuracy to be good and exceed some of the methods and models mentioned.

5 Conclusion and Future Work

5.1 Conclusion

This research project aimed to classify text essays accurately using different machine learning models. A consistent set of 20 features was used across all experiments for fairness and comparability. The results showed that the Gradient Boosting Classifier achieved the highest accuracy (0.70) due to its boosting technique. The model's accuracy (0.71) surpassed some systems and approached that of the multiple linear regression method (0.77). Overall, the research successfully explored various models and highlighted the importance of complexity, parameter tuning, and data handling for accurate classification. The Gradient Boosting Classifier emerged as a promising model for this task, showing competitive performance compared to other methods.

5.2 Future work

In order to enhance the accuracy and evaluation capabilities of the automated Arabic written evaluation system, there are many potential avenues for future work from using libraries of natural language processing (NLP) technologies that are specifically designed for Arabic language processing by developing them and thus obtaining a more accurate evaluation. Increasing the number of features can help improve the performance of the system through the use of the selection feature, which selects the most features that affect accuracy and thus focus on them. In addition to that, increasing the number of the data set, which will inevitably help improve accuracy, includes trying to distribute the data set into classification groups so that there is fairness in the evaluation.

References

[1] Mansour Alghamdi, Mohamed Alkanhal, Mohamed Al-Badrashiny, Abdulaziz Al-Qabbany, Ali Areshey, and Abdulaziz Alharbi. "A hybrid automatic scoring system for Arabic essays". *Ai Communications* 27.2 (2014), pp. 103–111.

[2] Saeda A Al Awaida, Bassam Al-Shargabi, and Thamer Al-Rousan. "AUTOMATED ARABIC ESSAY GRADING SYSTEM BASED ON F-SCORE AND ARABIC WORDNET". *Jordanian Journal of Computers and Information Technology* 5.3 (2019).

[3] Abeer Alqahtani and Amal Al-Saif. "Automated Arabic Essay Evaluation". *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*. 2020, pp. 181–190.

[4] Chun Then Lim, Chih How Bong, Wee Sian Wong, and Nung Kion Lee. "A comprehensive review of automated essay scoring (AES) research and development". *Pertanika Journal of Science & Technology* 29.3 (2021), pp. 1875–1899.

[5] Dadi Ramesh and Suresh Kumar Sanampudi. "Coherence Based Automatic Essay Scoring Using Sentence Embedding and Recurrent Neural Networks". *International Conference on Speech and Computer*. Springer. 2022, pp. 139–154.

[6] Jill Burstein, Martin Chodorow, and Claudia Leacock. "CriterionSM Online Essay Evaluation: An Application for Automated Evaluation of Student Essays." *IAAI*. 2003, pp. 3–10.

[7] Jason Griffith, Bill Strelioff, and James Schnaider. "The Hockley Index". *2012 Dallas, Texas, July 29-August 1, 2012*. American Society of Agricultural and Biological Engineers. 2012, p. 1.

[8] S Neelakandan and D Paulraj. "A gradient boosted decision tree-based sentiment classification of twitter data". *International Journal of Wavelets, Multiresolution and Information Processing* 18.04 (2020), p. 2050027.

[9] Dadi Ramesh and Suresh Kumar Sanampudi. "An Improved Approach for Automated Essay Scoring with LSTM and Word Embedding". *Evolution in Computational Intelligence*. Springer, 2022, pp. 35–41.

[10] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical methods for rates and proportions*. john wiley & sons, 2013.

[11] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. "Universal sentence encoder". *arXiv preprint arXiv:1803.11175* (2018).

[12] Meisam Moghadam and Niloufar Jafarpour. "A Survey of Part of Speech Tagging of Latin and non-Latin Script Languages: A more vivid view on Persian". *LANGUAGE ART* 6.1 (2021), pp. 75–90.

[13] Abeer Alqahtani and Amal Alsaif. "Automatic evaluation for Arabic essays: a rule-based system". *2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE. 2019, pp. 1–7.

6 Appendix

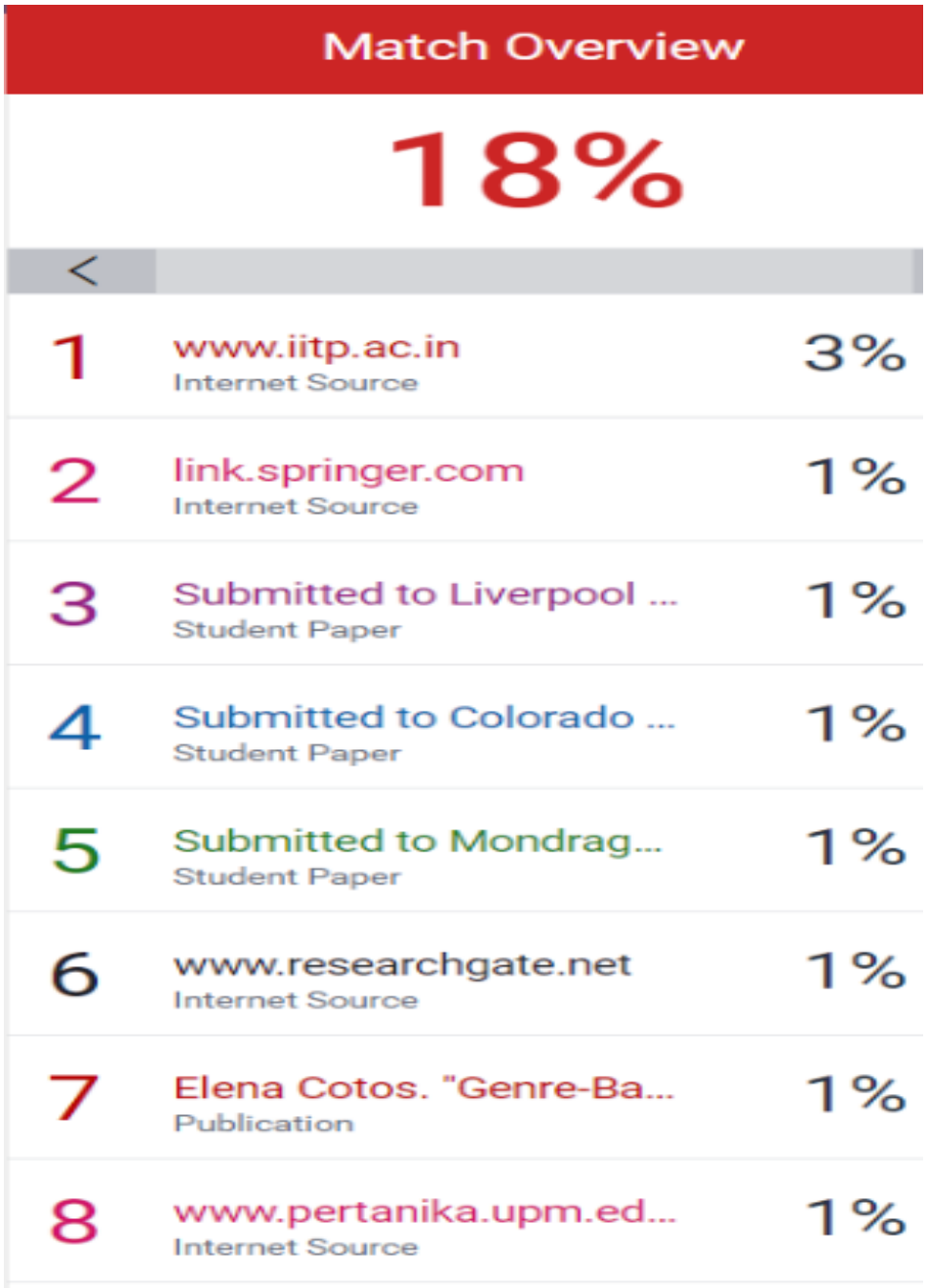

Figure 6.1: Copy ratio check