MEU جامعة الشرق الأوسط
MIDDLE EAST UNIVERSITY
Amman - Jordan

Toward an Arabic Essay Grading Benchmark for Machine Learning

بناء معيار لتصحيح الأسئلة المقالية في اللغة العربية لتعلم الآلة

Prepared by

Rawan Abed Al-Haleem Alzyadat

Supervisor

Dr. Bassam Al-Shargabi

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of master’s in computer science

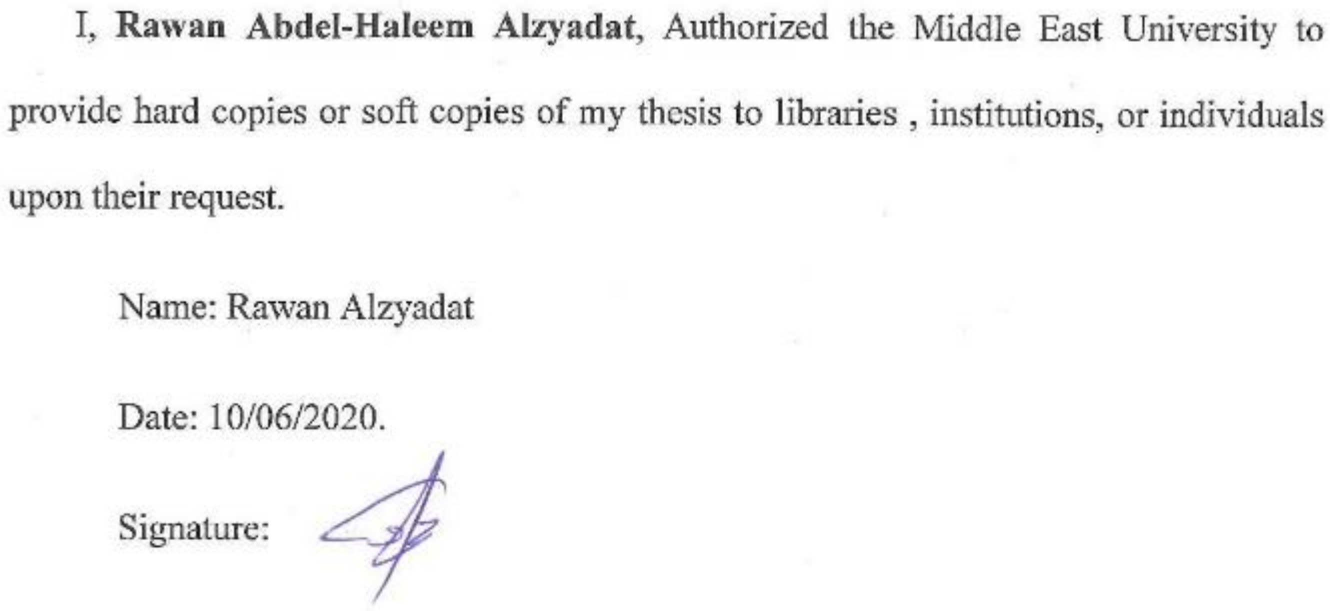
Department of Computer Science

Faculty of Information Technology

Middle East University

June, 2020

Authorization



I, **Rawan Abdel-Haleem Alzyadat**, Authorized the Middle East University to provide hard copies or soft copies of my thesis to libraries , institutions, or individuals upon their request.

Name: Rawan Alzyadat

Date: 10/06/2020.

Signature:

Thesis Committee Decision

This is to certify that the thesis entitle "**Toward an Arabic Essay Grading Benchmark for Machine Learning**" was successfully defended and provide on 27/5/2020.

Examination Committee Members	**Signature**

Dr. Bassam Al-Shargabi (Supervisor/chairman) ....

Associate Professor, Department of Computer Sconce

Middle East University(MEU)

Dr. Ahmed Al-homouz (Internal Examiner) ....

Associate Professor, Department of Computer Information System

Middle East University(MEU)

Prof. Nidal Turab (External Examiner) ....

Professor, Department of Computer Science

Al-Ahliyya Amman University(AAU)

Acknowledgment

Many thanks are submitted first and foremost to Allah who gave me the strength and ability to complete this study.

Words can never express my hearty thanks to the supervisor, Prof. Bassam AL-Shargabi for giving me invaluable guidance, insights, moral support, and the direction throughout the whole process of completing this thesis. I will never forget what he had done for my sake.

I would like to express my great thanks and appreciation to all my professors at the Faculty of Information Technology, Middle East University for their continuous help and assistance in understanding research by providing me with materials to be used in this study. Their ideas and discussions helped me in developing new ideas for this writing.

Most importantly I dedicate this study and effort to my parents and my husband for their trust in my choices, morally and financially providing me the invaluable support making it possible to go after my dreams.

Last but not least to all my friends who never give up in giving me support information and assistance in completing this study. Award of thanks also extends to those who have indirectly provided comments and helpful suggestions.

I have always believed in this man Mahatma Ghandi and his wisdom-" In doing something do it with love or never do it at all ". Through the process of understanding and learning the research process.

Rawan Al-zyadat

The Researcher

Dedication

This thesis is dedicated to my father, Abdel Al-Haleem Alzyadat, for believing in me and for his constant support in accomplishing this thesis.

I would like also to dedicate it to my husband, Wesam Alzyadat, who helped me psychologically and financially and who bore my troubles during my journey into achieving my childhood dream.

To my mom, sister, and brothers for their limitless love and support. I will never forget what they had done for me. I wish to express my heartily

Gratitude and thankfulness to my daughter for being the primary motivation for struggling every day to finish this study. I want her to be proud of me in the near future.

To all people who encouraged me, I am grateful to you.

Rawan Al-zyadat

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Meaning
AEG	Automatic Essay Grading
AI	Artificial Intelligence
ARFF	Attribute-Relation File Format
ASAP	Automated Student Assessment Prize
CNN	Convolutional Neural Networks
CSV	Comma Separated Values
DL	Deep Learning
E-rater	Electronic Essay Rater
GLM	General Linear Model
IEA	Intelligent Essay Assessor
KNN	K-Nearest Neighbors
ML	Machine Learning
NB	Naive Bayes
NLP	Natural language processing
PEG	Project Essay Grader
RNN	Recurrent Neural Networks
SVM	Third Support for the vector machine
WEKA	Waikato Environment for Knowledge Analysis

Toward an Arabic Essay Grading Benchmark for Machine Learning

Prepared By:

Rawan Abed Alhaleem Alzyadat

Supervised By:

Dr. Bassam Al-Shargabi

Abstract

Using Automatic Essay Grading (AEG) is seen as a vital tool, as it provides a lot of advantages like getting the marks without human bias, quick and safe effort. Nowadays, the majority of grading systems have become dependent on artificial intelligence such as machine learning. As a result, most of AEG systems utilized machine learning to grade essays based on prepared dataset for training and validation.

For the English essays there is an Automated Student Assessment Prize (ASAP) dataset for grading essays using machine learning but there are no Arabic essays grading datasets for machine learning. Therefore, this thesis is an attempt to collect and establish a dataset for Arabic essay grading for machine learning. In this thesis, the established dataset or benchmark contains essay questions with their graded model answers for a various topic across most all different school levels. The collected dataset was tested and evaluated using the best-known classification algorithms such as Naive Bayes, decision tree, and meta classifier. The experimental result shows that dataset is suitable for the machine learning algorithms, where the classifiers performance results shows a 79%, 81%, 86% as accuracy based on established dataset.

Keywords: Automatic Essay Grading, Classification, Naive Bayes, Decision Tree.

بناء معيار لتصحيح الأسئلة المقالية في اللغة العربية لتعلم الآلة

إعداد

روان عبد الحليم الزيادات

إشراف

الدكتور بسام الشرجبي

المٌلخص

يعد التصحيح التلقائي للأسئلة المقالية أداة حيوية، لأنه يوفر الكثير من المزايا مثل الحصول على العلامة دون تحيز بشري، وجهد سريع وآمن. في الوقت الحاضر، أصبحت غالبية أنظمة التصحيح التلقائي للاسئلة تعتمد على الذكاء الاصطناعي مثل التعلم الآلي. ونتيجة لذلك، معظم أنظمة التصحيح التلقائي تعتمد حاليا على التعلم الآلي للتصحيح التلقائي للأسئلة المقالية، حيث يتم تصحيح الاسئلة اعتمادا على مجموعة من الاجابات النموذجية معده مسبقا لاستخدامها في التدريب والتحقق من صحتها.

بالنسبة لمقالات اللغة الإنجليزية، توجد مجموعة بيانات ASAP لتصنيف المقالات باستخدام التعلم الآلي ولكن لا توجد مجموعات بيانات للتصحيح التلقائي للأسئلة المقالية لتعلم الآلة. لذلك، هذه الرسالة هي محاولة لجمع وإنشاء مجموعة بيانات لتصحيح الأسئلة المقالية لتعلم الآلة. تحتوي هذه الرسالة على مجموعة البيانات المعتمدة على الاسئلة المقالية مع إجابات نموذجية متدرجة لمواضيع مختلفة ولمستويات مختلفة.

تم اختبار وتقييم مجموعة البيانات التي تم جمعها باستخدام خوارزميات التصنيف الأكثر شهرة لتصحيح الاسئلة المقالية تلقائيا باللغة العربية مثل (Naive Bayes، decision tree ، meta classifier) . تُظهر النتيجة التجريبية أن مجموعة البيانات مناسبة لخوارزميات التعلم الآلي، حيث تُظهر نتائج دقة الخوارزميات على مجموعة البيانات كالتالي 79٪ و81٪ و86٪.

الكلمات المفتاحية: التصحيح التلقائي للأسئلة المقالية (AEG)، التصنيف ، خوارزمية Naive Bayes ، خوارزمية شجرة القرار decision tree .

Chapter One

Introduction

Chapter One

Introduction

1.1 Introduction

As the number of schools, colleges and universities are increasing at a rapid pace, so the number of admissions of students. To providing quality education to the students, it is essential to evaluate the student's result unbiased to infer the actual results. To avoid this particular problem, the Automatic Essay Grading (AEG) is used, where human bias, quick and save efforts are handled properly with the help of computer algorithms. AEG is actively used in developed countries that are using English as their chosen language. But, implementing the same with the Arabic language is not possible because of the lack of an Arabic dataset.

The main reason behind this lack of dataset is the core structure of Arabic literals that increases the challenges. There are so many words with their synonyms that also have some words which do not have any meaning without combining it to their source. And interestingly, some of the words used in Arabic essays are not even included in the Arabic Language Glossary (Abdulaziz, Mahmoud, &Magdi, 2018).

While comparing the English AEG with the Arabic AEG, the main problem arises while referring to the sample dataset for authentication and verifying the same. In English AEG, a 13,000-essay open-source dataset is available for assessment which is the main source of their grading system such as the Automated Student Assessment Prize (ASAP). In Arabic AEG, there is not much progress in this field as there are no such referencing datasets is provided for the research to upgrade the progress and accuracy of the required AEG (Mathias & Bhattacharyya, 2018).

In this thesis collected and establish a dataset for Arabic essay grading, that it contains a set of question and each question has a set of answers, this benchmark prepared to be suitable for machine learning algorithms, where classification algorithms are used to validate and test the benchmark.

1.2 Definition of Study

1.2.1 Automatic Essay Grading

In order to grade the essays in the past, there was only one method available that is evaluating the essay line by line, giving remarks and marks. After marking the paper, the teacher has to total the marks and give them an appropriate grade. This process is time-consuming and may be biased for some students. There are chances that manual grading may be inefficient. (Stanojevic,2019)

As the above system consumes time, the AEG is more efficient, effortless, and more accurate in any sense. This is very convenient and helpful for both the teachers and the students.

Basically, there are three types of AEG for grading is available. One of them is Project Essay Grader (PEG). It was first used by the Ellis Page in 1966. The main focus of this thesis is to evaluate the score by considering the quality of the students' essays. To implement the proposed system, Page et al use feature extraction such as paper length, conjunction, word length, and punctuation marks. After getting the desired output from the former method, this is taken as the input for the regression model for an essay assessment. PEG provides quite effective results with 87% accuracy (Rudner ,& Gagne, 2000).

The second type of AEG is The Intelligent Essay Assessor (IEA), which is suggested by the Landauer and Foltz in 1997. The basic idea behind the research is to evaluate the correlation between the humans and the IEA scores. The score is calculated on the basis of similarity after converting each essay into vectors and identifies the words that are more frequent than the calculated similarity. If the correlation between human and IEA core is high, then it is considered as a good and acceptable result. (Rudner, & Gagne, 2000)

The third type of AEG is the Electronic Essay Rater (E-rater), which is created by Jill Burstein in 1998 to analyze the features. E-rater has 60 various features that represent the data in the form of rows and columns. This is further vectorized and the word frequency is calculated to compute the similarity between the essays. To implement this, a regression model is used, to predict the score of the students (Rudner, & Gagne, 2000).

1.2.2 Benchmark

Benchmark includes the collection of datasets in tabular form. This means that the data set stored in the combination of rows and columns on which the Machine Learning or ML models heavily depend. This helps the ML models to measure the accuracy and the back-testing of future technologies (Sinka, & Corne, 2002).

Firstly, the benchmark is the set of standards that must be followed to be used as the point of reference. This is usually done to compare similar models and evaluate the accuracy, quality, and effectiveness (Ramsundar et al, 2018). Usually, the benchmark follows the three important stages to make it effective and accountable such as data collection, structuring of collected data, and validating the data using performance measures (Ramsundar et al, 2018).

Among the three stages of benchmarking, data collection is the most important and critical part. As the model is totally dependent on the gathered data, so verifying the source and cleaning of collected data is the most critical part of this phase. There are some other important aspects that are also considered in data gathering like the selection of sample data strategy i.e. sampling method and sampling size, division of data.

There are some questions that must be answered before starting the data collection phase like the purpose of the benchmark, function of the benchmark. Second, which tools are going to be used to build the benchmark? Third, how do we collect the desired data and which methods are we going to follow to fulfill this. Fourth, identification and verification of the collected data are performed.

The step of the benchmark is to build the dataset. This includes defining the structure of the collected data that must be analyzed and organized by a specific tool and defined method. To ensure credibility, the dataset must be perfect and must include important information. The validation of the data must be ensured before measuring the performance and the analysis.

In this thesis, we established and designed a dataset for Arabic AEG by using numerous sets of questions, where each question has typically three answers. Datasets may vary in regards to topics such as Sciences, Geographically, Islamic, Computer, and others.

1.2.3 Machine learning

First, it must be bear in mind that there is a difference between Artificial Intelligence (AI), Machine Learning ML), and Deep Learning (DL). The difference is not clear to everyone. The definition of AI is a machine capable of performing a certain task of human intelligence to perform. In AI systems generally, there are at some of the features consist of: planning, learning, thinking, problem-solving, representation of knowledge, perception, movement, and creativity.

The other terminologies ML and DL are already the talks of academicians. Basically, ML is just a mean to achieve artificial intelligence and is the ability to learn without explicit programming, this thesis will focus on ML and specially supervised type. DL is a method of machine learning, which is based on the structure and functions of the brain, the relationship between many neurons. Neural networks are an algorithm that mimics the biological structure of the brain. The Figure 1.1 shows the relationship between AI, ML, and DL (Jakhar,& Kaur 2020).

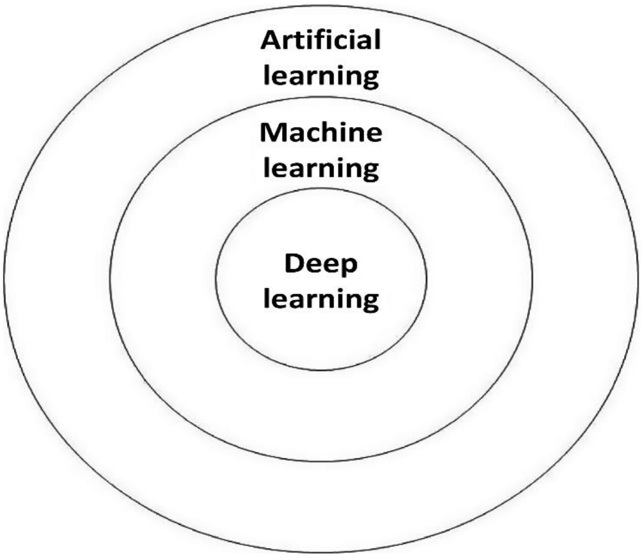


Figure1.1: The relationship between AI, ML, and DL (Jakhar, & Kaur 2020).

ML by definition means to create algorithms that can receive input data and use statistical analysis to predict outputs within an acceptable range. it is generally divided into three categories: supervised, unsupervised and semi-supervised learning.

In the first type, the supervised with unlabeled training dataset is very simple, which means the existence of data entered (x) so we have in the output data (Y) and the algorithm takes this data, so that when came data of type (X) new machine provides me with data (Y) Based on their training (Kotsiantis, Zaharakis & Pintelas, 2007). In the second type unsupervised (with only labeled training dataset) give the computer data (X), but you do not know what the output (Y) and ask the machine to give you the output (Y), it is given only data and it calculates to give you the output, Figure 1.2 shows types of machine learning (Society, 2017).

In the third type, the semi-supervised learning in the dataset contains both labeled and unlabeled data. Where most of the input is self-evidently unlabeled.

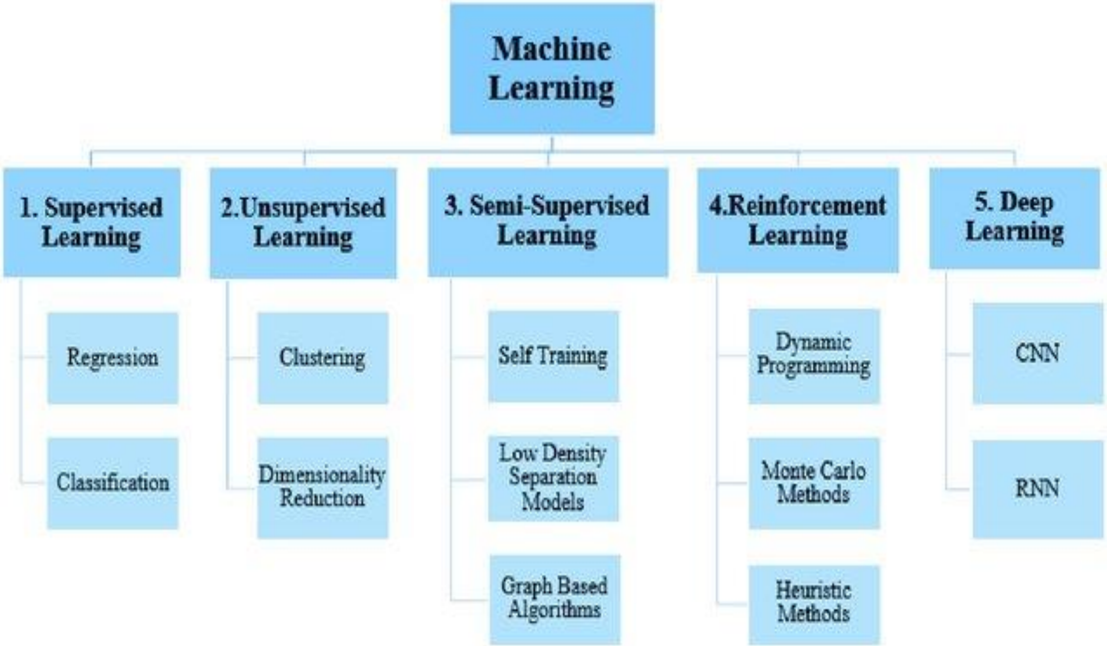


Figure 1.2: Types of Machine Learning (Nassif et al.,2019).

1.3 Problem Statement

Automated essay grading system using machine learning helps the teachers to lower their burden of grading the papers or essays. But the main challenge occurs while implementing it for the Arabic language, where lack of sufficient dataset creates the major barrier for the researchers. This thesis handles the problem of the Arabic Essay Grading Dataset that will be used by the machine learning models to predict more accurate results.

This thesis aims to develop a centralized Arabic essay benchmark for machine learning, where the benchmark consists of a large set of questions and answers.

1.4 Research Questions

To attempt addressing the limitations discussed in the previous section, the following specific questions are posed:

1. Will maintaining an Arabic essay benchmark or dataset will improve automating Arabic essays grading using machine learning?
2. How can machine learning help in improving automating Arabic essay grading?

1.5 Goal and Objectives

The main objectives of this thesis can be summarized as follows:

1. Proposing an Arabic essay grading benchmark using machine learning.

2. Apply multiple machine learning algorithms in order to compare the accuracy based on the developed benchmark along with other measures.

1.6 Motivation

As the number of students increased in the universities, teachers are devoting their time to grading the students' essays rather than improving the skills. The automated essay grading system is very useful for the teachers to automatically grade the paper unbiased and hectic free. This free time will allow the teacher to focus on making students' abilities to improve rather than debugging the exam copies one at a time. This also increases the effectiveness and efficient grading system that can be trusted by the students and the teachers.

As there is no such benchmark is available for the Arabic language as a present for the other languages like ASAP. This helps to improve the efficiency and quality. By creating an Arabic benchmark for everyone to improve the efficiency of the machine learning model to solve the existing set of problems.

1.7 Contribution and Significance of Research

The proposed Arabic Benchmark for Automated Essay grading dataset by using machine learning will lead to the development of accurate Arabic AEG. This helps the researchers to improve the grading system and alternatively, provides the more free time to the teachers to focus on their own skills to improve so that they can help the students to get the enhanced desires skillset eventually. The proposed Arabic dataset has also led the foundation for further research.

1.8 Scope and Limitations of the Study

The scope of the study is to design and implement the Arabic essay grading benchmark for machine learning. Proposed dataset is suitable for the Automate Arabic essay grading based on machine learning. This study will be limited only for the grading essays written in the Arabic language.

1.9 Thesis outline

This chapter provides an introduction about AEG, describes the benefits of the student and the teacher when using it. Also, explains Arabic essays that have different natures than English essays. Additionally, the research problem, research questions, goal and objectives, motivation, contribution, and significance of the research, and limitations of the study are also discussed. The rest of this thesis is organized as follows:

Chapter Two presents the background essay grading system, machine learning, and most recent related works regarding the topic of this thesis.

Chapter Three discusses the proposed methodology that was adopted in work on a collected dataset for Arabic Essay Grading and explains the method of data collection, and proposed design dataset.

Chapter Four presents the implementation of the proposed descriptor. The results and their effectiveness are also discussed in this chapter.

Chapter Five will give a general summary of the thesis, summarizes the research findings and future works.

Chapter Two

Background and Literature Review

Chapter Two

Background and Literature Review

2.1 Introduction

This chapter presents the background of AEG, Machine Learning Algorithms, and Benchmark. Describes how to develop essay grading, what they are algorithms used. This study for the dataset used the best-known classification algorithms (Naive Bayes, decision tree, and meta classification). This chapter also presents the advantages and disadvantages of the algorithms used to test the dataset. The most recent related works to the concept of AAEG and latest used datasets or benchmark is also presented in this chapter.

2.2 Background of Automated Grading System

The automated grading system is an interesting topic in education and very important because everything is used here is toward technology, in general, the tests have a lot of techniques like (multiple choice, matching, true/false, short answers questions, open questions, and other techniques. The automated grading system is very helpful and important as it helps to improve efficiency, quality, and save time in correction exams. In the past, using manual essay grading was very difficult since it takes a lot of time, the possibility of teacher bias and providing incorrect grading for students this way is a highly inefficient, now after increased students in school and university must uses automatic essay grading it's very useful for teachers and students, Through the use of AI it provides more efficient essays grading.

Began researches to get system automatic essay grading beginning work of Project Essay Grade (PEG) in 1960 it Use General Linear Model (GLM) to predict the outcome of the essay, analysed samples of the essays contain 495 and 599 essays, the result accuracy was 87 % it was high of accuracy, this the first project of automatic essay grading (Batten, 1994).

After that new system came, the students were able to type directly the computer and use binary classification, the essay divided two way "good" or "bad", used classifier worked for similarity weight between training essays and testing essays (Larkey, 1998).

Then came the new work to show grading in short essay answers by Matching in model answer with student answer and show the grading ,that mainly used three algorithms together, the system result 82% correlated with the human score (Ali & Mohd, 2013).

2.2.1 Types of Automated Grading System

There are Three main types of (AEG) that are used to grading Essays, the first type was the Project Essay Grader (PEG) that was established by Ellis Page in 1966, measured the score by the quality of essay by using a regression model and feature extraction like (word length, paper length, conjunction, and punctuation marks), (PEG) is get good results and apparent reliability arrive at 87%.

PEG works in two parts the first training essays, in this part make analyses of the essay and calculates over 500 features of writing like diction, grammar, and so on. second after calculated all features PEG used to prediction essay grading.The second type of Project Essay, Intelligent Essay Assessor (IEA), the first one is suggested for the essay scoring by Landauer and Foltz in 1997. The principle works to identify which the essays are most similar to the new essays, each essay converts to vectors and identifies where the words the most frequent then calculated the similarity, focus on the content, the system based on LSA to evaluate essays. The IEA used the technique to analyse essays compare similarly to the essay with other essays and quality of content.

Training phase needed 100 essays smaller than other types of AEG after that to predict essay grade. The result showed high correlations between human and IEA scores it is a good result (Semire, 2006).

Electronic Essay Rater (E-rater)create by Jill Burstein in 1998,work to analysis features, e-rater uses 60 various features, representing the data from rows and columns to vectors after that is use the similarity between essays and make calculation of words frequency in each essay by using a regression model to predict the score of students (Rudner & Gagne ,2000).

Other systems used machine learning to grade student’s essays such as: using linear regression to predicted grades, matchings between predicted grades and human grades. Here, the final result showed the ability to predict student’s score it is good. They data used from kaggel.com, by William and Hewlett it contained 1300 essays, each one length between (150 to 550 words), the data include 8 groups of essays (Manvi, Mishel, & Ashwin, 2012). Another advanced method used deep learning methods such as the Long Short Term Memory (LSTM) networks to show the meaning of texts, able to get very good results. The dataset used from Kaggle included 12.976 essays each one length between (150 to 550 words), divided to 80% training/validation, and 20% for testing (Dimitrios, Helen, & Marek, 2016).

2.3 Machine Learning

ML is part of the artificial intelligence which made the computer have the ability to learn and to think like the humans, automatically learning without human interference and developing learning over time. The aim of machine learning was getting computer interactions with the real world as can access the data required and used to learning (Chen & Liu 2018).

2.3.1 Types of Machine Learning Algorithms

In general, there are four types of machine learning algorithms: the first, supervised learning which contained training labelled data that predicted the outcome of new data, the system allowed providing a target for any new data after good training if it predicted incorrectly. The last step made the comparison between the output data and the correct data and discovered the error and corrects the model. Secondly, the unsupervised learning dealing with unlabeled and hidden structure data, used to predict data outcome without guidance. Third, semi-supervised learning can put between supervised and unsupervised machine learning, a lot of data labelled and some unlabeled so it closer to supervised learning. Fourth, reinforcement machine learning this type is being supervised in general just the principles, and the prediction conduct self-learn by interface with its environment (Ayodele, 2010).

In this thesis, we are focusing more on classification algorithms to be used to validate the established benchmark or dataset, where the essays grading will be dealt as classification problem.

2.3.2 Classification Algorithms

It is a technique by which you can train a dataset on certain conditions that can be used to identify each target group and to predict the target group. The simplest example is a binary classification. The types of classification algorithms as follows:

2.3.2.1 Linear Classifiers

It contains logistic regression and Naive Bayes classifier. Logistic regression is the simplest example is a binary classification where data is either in the first or the second set, that is often used to know one of the properties of the thing we are studying example (0/1, yes/no, true/false, and Male/Female).

Naive Bayes classifier define “Naive” means the very simple algorithm, “Bayes” is referred to Bayes theorem, it is one of the most famous methods of learning machine, where characterized by the speed in processing and efficiency in the prediction; that it took on the principle of Independence Assumptions so that the relationship between all attributes, features are seen as independent of each other.

Naive Bayes was used in many systems for example in identifying spam messages, in the classification of the documents to predict the type of document (politics, sport, technology) text classification, to recognize points of view (negative, positive, optimistic), and other usages like face recognition in pictures. It gave very great results when using it in texts (Rahul, 2017), the Bayes Theorem it is calculated as in equation 1:

$$\mathrm{P}(c|d) = \frac{P(C)\,P(d|c)}{P(d)} \qquad \text{(1)}$$

where C = {c1, c2, ………, cm} possible classes. D = {d1, d2, ………., dn} domain of words. P(d): probability is a constant for the dataset size (Liu, Blasch, Shen, & Chen, 2013).

- **Advantage**: Simple that can be implemented easily and effectively, uses logistic regression as a standard and then uses more complex and more difficult algorithms (Niklas, 2018).
- **Disadvantages**: of the NB classifier does not work well in the numerical dataset and cannot be solved any problem non-linear(Naresh, 2019).

2.3.2.2 Decision Trees

A decision tree or DT is a subset of machine learning and also the most commonly used ML model. DT usually divides the dataset into a smaller and manageable dataset on the basis of a predefined attribute and places it in the tree branches for the simplicity and organizational purpose. The most popular types of DT are the random tree, regression tree, iterative dichotomized, and j48, which we are using in our thesis.

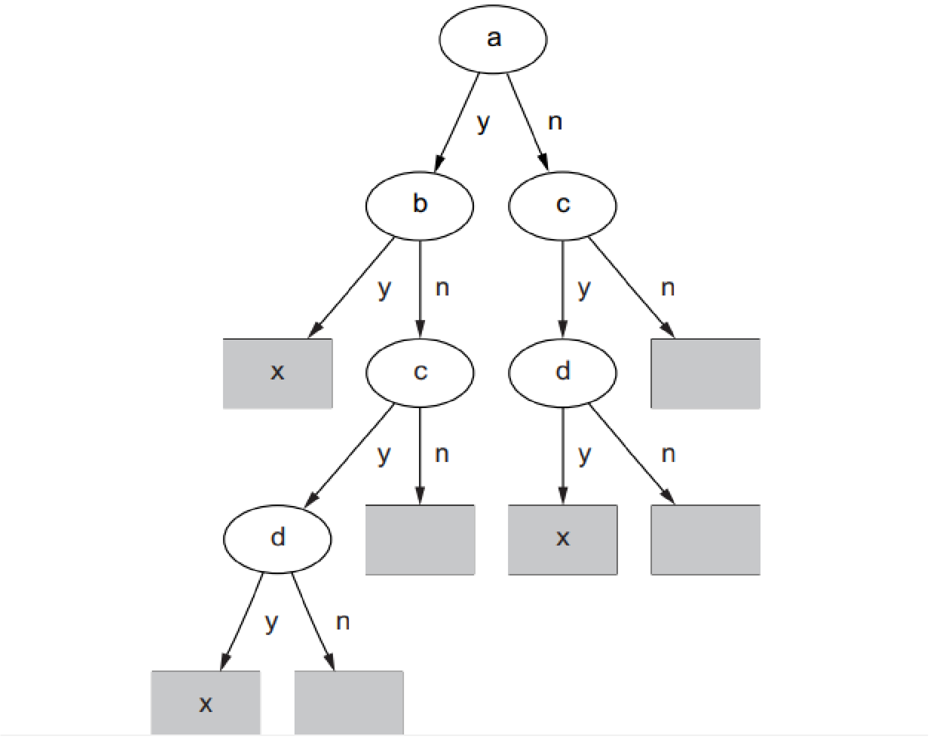


Figure 2.1: Decision tree for a simple disjunction. (Witten, Frank, & Hall, 2005).

Advantage:

1. Simple algorithm,
2. Very high accuracy,
3. The training phase is very fast and
4. Easy to understand and Implement.

Disadvantages:

1. The testing takes a lot of time

2. If there is any mistake in the beginning of the tree then the sub-trees also inherit the same mistake with the same accuracy.

2.3.2.3 Meta-Classifier

Meta classifier is efficient when the lack of ideal machine learning failed to persist. It is used to implement in the workplaces, just to improve the accuracy. The hybrid algorithms produce more accurate results than the former once (Tastle & Terzi 2013).

Usually Meta-Classifier splits the training dataset into two levels and use the first level output as the input for the second level. This process was repeated as illustrated in the Figure 2.2 (Nguyen & other, 2019). Figure 2.2 shows that the improved performance of the collected dataset using two types of decision tree such as Forest algorithms and J48.

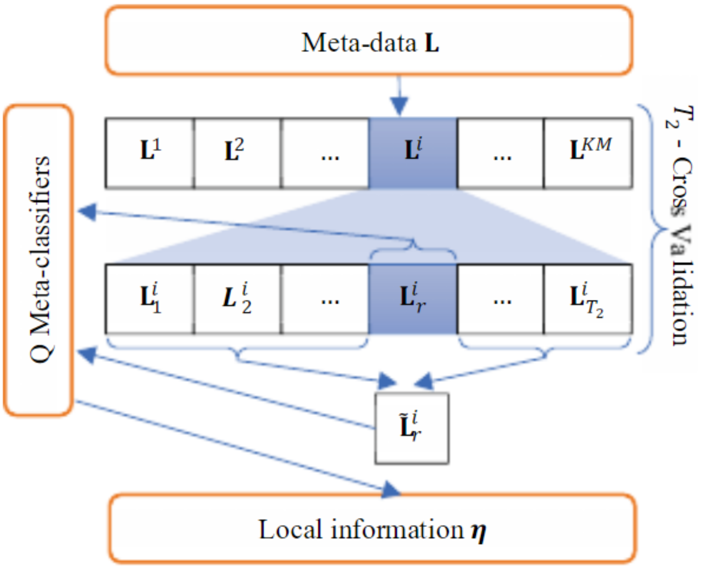


Figure 2.2: The Process of Meta-Classifier.

Advantage:

1. It is the most important machine learning algorithms,
2. Considered reliable classification,
3. high efficiency, and accuracy at the same time have little time and space complexity.

2.3.3 Steps of Machine Learning

The implementation of machine learning process is divided into five major steps such as data collection, data cleaning or pre-processing, training, testing and improving the model as illustrated in figure 2.3.

First step of machine leaning includes the data collection from the different sources like Kaggle,GitHub or online paid or open-source repositories. When structuring the dataset in a pre-defined format, there are some guidelines that must be followed for the authentication purpose. Some of them are: sources identification, mode of data sampling strategy (sampling method and sampling size), split portion of data for training and testing (Matthew, 2018).Second step of machine learning is very critical and very important as the whole model depends on this process. This phase usually provides the suitable, structured and desired dataset to the model. Pre-processing of the gathered data includes the scaling features, data preparation, data selection, removing duplicate entries, filtering unusual data. Next step is used to decide the model to be implemented on the dataset. Model selection is dependent on the type of dataset. Model accuracy, complexity of model, the time taken by the building process, training & testing and effectiveness of the model are the key factors that are considered while selecting the model.

Training and testing phase of machine learning is simply used to divide the whole dataset into 80-20 or 75-25 proportion respectively. This means that the 75% of dataset is reserved for the training of the model and 25% of dataset is reserved for the testing of the trained model. This helps the model in predicting the target accurately (Matthew, 2018).The last of machine learning is the evaluation phase which is implemented by using the testing process. As the testing data is independent of the training data, this is used to predict the actual results and to improve the performance of the selected model.

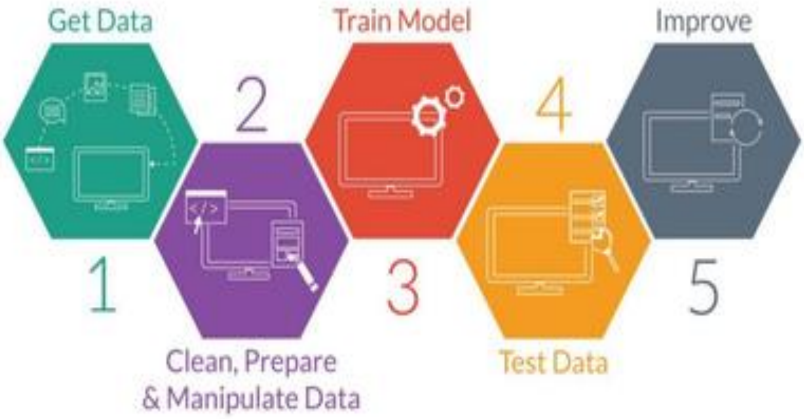


Figure 2.3: The Steps of Machine Learning (Aileen, 2018).

2.4 Related Works

This chapter is intended to illustrate literature review section on Arabic AEG Benchmark for machine learning. The literature reviewed in this section outlines the possible methods, techniques, and technologies used in AEGs for machine learning with their predicted results and finally display the summary of this chapter. In this chapter, some of the developed studies will be discussed

In Saad, (2010), the authors provided a study of the impact of Text Pre-processing and Arabic text analysis, the analysis tools used open-source machine learning tools: WEKA. while The Arabic analysis used Arabic stemmer (stemming and light stemming), and used algorithms to classifiers Naive Bayes (NB) and Support Vector

Machines (SVM), the dataset collected from multiple Arabic websites like CNN and BBC Arabic website this is the largest Arabic text dataset; 33 different Arabic text and contain 18M words. Takes the text dataset and calculates the most word frequently s bag-of-words this called feature extraction after that detect what the text takes about (sport, business, and education), and work in natural Language Processing (NLP).

In Alsaleem, (2011) the authors proposed a supervised machine learning to classify Arabic articles for Saudi Newspapers. Compare the results of two algorithms, Support Vector Machine algorithm (SVM) and Naïve Bayesian (NB). Arabic documents different lengths and different categories such as (economy, public, social technology, sports) Pre-processing are very important in English and Arabic, but in Arabic is different and more difficult because of the complexity of the language. They divided data into groups and represented the data in a way that fit the algorithms used.

In Gomaa ,& Fahmy,(2014) the authors provided system Arabic essay scoring using Arabic text to English translations because of the dearth of essays processing in Arabic, it the first benchmark dataset in Arabic that Includes 610 short answers for the students and their English translations, includes four types of question (Define, what, why and explain) it works to translated text in Arabic to English after that uses to k-mean clustering and similarity text to give a result. To check Arabic essays using supervised techniques linear regression and simple linear regression by using WEKA tools.

Mezher, & Omar, (2016) described for the model they suggest contains two methods for (AEG) Semantic Analysis and syntactic features. the syntactic feature makes the accuracy of AEG more efficient; the dataset Includes 61 question and each question have 10 answers, the number of total answers 610. To check Arabic essays using techniques Latent Semantic Analysis (LSA), using Term Frequency-Inverse Document

Frequency (TF-IDF) and the result by using LSA is 0.745 compared by other techniques.

Al-Jouiea, & Azmia, (2017) proposed a system in the Saudi National Centre that conducts a standard exam for any student wishing to enroll in any of the national university. The General Ability Test (GAT) There is a large number of student applicants. Working in the field of Natural Language Processing (NLP) and latent semantic analysis (LSA), was applied to evaluate children's essays in schools, which are in Arabic, and was tested on three hundred essays in different areas and was interested in the consistency of words, spelling, style and achieved high results up to 77%.

Abdulaziz, Mahmoud, & Magdi, (2018) proposed an Arabic essay grading (AEG), where questions were divided two types: long and short answer. The dataset contained 21 questions 210 short answers used text similarity algorithms to show the students grading. The system based on text similarity and compare student answers with a model answer, it is working to take answers make some steps to prepare data before training like (data pre-processing, Arabic WordNet, and feature detected applying text similarity algorithms to show the score of the student’s answer.

Alawaida, Al-Shargabi, & Al-Rousan, (2019) they proposed the Automated Arabic essay grading model by using two techniques: the first using support vector machine (SVM) and text similarity algorithms. They used a dataset contains 40 questions and 120 answers, where they used F-score to extract features from student answer and ideal graded answers and apply cosine similarity measure to score student answer.

2.5 Summary

In summary, the purpose of this study is to establish a large dataset for the Arabic AEG. As of today, there is no such Arabic dataset available for the researchers to pursue their research in this field. There are many small but unstructured datasets are available. These small datasets are not as reliable as the number of sampling data is low. So, we need to classify them according to our study and build a centralized structure dataset. Existing datasets contain a maximum of 612 short questions and to use them researcher has to convert it into English. This does not solve the problem of Arabic AEG. Our study provides the 3000 short answers and 35000 words of diversity to make it large enough to produce accurate and reliable results.

Table 2.1 is the structure and organized form of related work in the field of Arabic automated essay grading system for machine learning. The table explains the used datasets and the machine learning algorithms by the researchers. The study of the above table clarifies the research motives, tools, and findings in an organized manner.

Table 2.1: Summary of Related Work.

No.	Authors	Description
1.	Saad,(2010)	They used open-source machine learning tools: WEKA. The Arabic analysis used Arabic stemmer (stemming and light stemming), used algorithms to classifiers Naive Bayes (NB), and Support Vector Machines (SVM). This research used algorithms to classifiers (NB) in WEKA tools.

No.	Authors	Description
2.	Alsaleem, (2011)	The authors describe supervised learning automatically classified in Arabic articles for Saudi Newspapers. Using two algorithms Support Vector Machine algorithm (SVM) and Naïve Bayesian (NB) and compare algorithms results. This research used algorithms for classifiers (NB) and Pre-Processing on Arabic text.
3.	Gomaa, &_Fahmy , (2014)	It provides system Arabic essay scoring using Arabic text to English translations because of the dearth of essays processing in Arabic, it the first benchmark dataset in Arabic by using WEKA. This research creates a dataset without English translation by using WEKA.
4.	Mezher ,& Omar, (2016)	They suggested the two methods for (AEG) Semantic Analysis and syntactic features check Arabic essays using techniques like Latent Semantic Analysis (LSA), using Term Frequency-Inverse Document Frequency (TF-IDF) and compared by other techniques.
5.	Al-Jouiea,& Azmia,(2017)	The authors described natural language processing (NLP), and latent semantic analysis (LSA), and using

No.	Authors	Description
		feature extraction like a bag of words was interested in the consistency of words, spelling, and style applied for training essays after that evaluate children's essays in schools which are in Arabic, then tested. This research used feature extraction for training essays
6.	Shehab, Faroun,& Rashad,(2018)	The dataset contained 21 questions and 210 short answers that use text-like algorithms to show student grades. The system was based on text similarity and performs some steps to prepare the data before training such as data pre-processing of the Arabic WordNet, and the feature discovered when applying text similarity algorithms to show the degree of answer student.
7.	Al-awaida, Al-Shargabi, &Al-Rousan, (2019)	The dataset contained 40 questions and 120 answers, the system works to take answers after pre-processing, compare student answers with a model answer by two techniques (SVM), text Similarity and show the Student score. This research used the pre-processing of the dataset.

Chapter Three

Methodology and the Proposed Model

Chapter Three

Methodology and the Proposed Model

3.1 Introduction

In this chapter, the methodology that followed in work on a collected dataset for Arabic Essay Grading and explains the method of data collection, and designed dataset. Describe the processes to organize the data well and Describe the steps of data collected and how validated and evaluated using three machine learning algorithms (Naive Bayes, Decision tree, and meta classifier) it works to show the best results.

3.2 Proposed Methodology

The methodology used in this thesis is shown in Figure 3.1, are consists of steps summarized as follows:

1. Data collection, must be gathering clear data from the original place, identify sources, select of the dataset.
2. Data structure, data building in different fields (Sciences, Geographically, Islamic culture, Computer, and other), each question has 3 typical answers.
3. Data pre-processing and analysis. (Transformation data as required by the application
4. Dataset is divided into two parts: Training, and Testing.
5. Training dataset using a set of machine learning algorithms.
6. Validate dataset using machine learning algorithm, choose appropriate algorithm to give the best results and compare them with other algorithms.

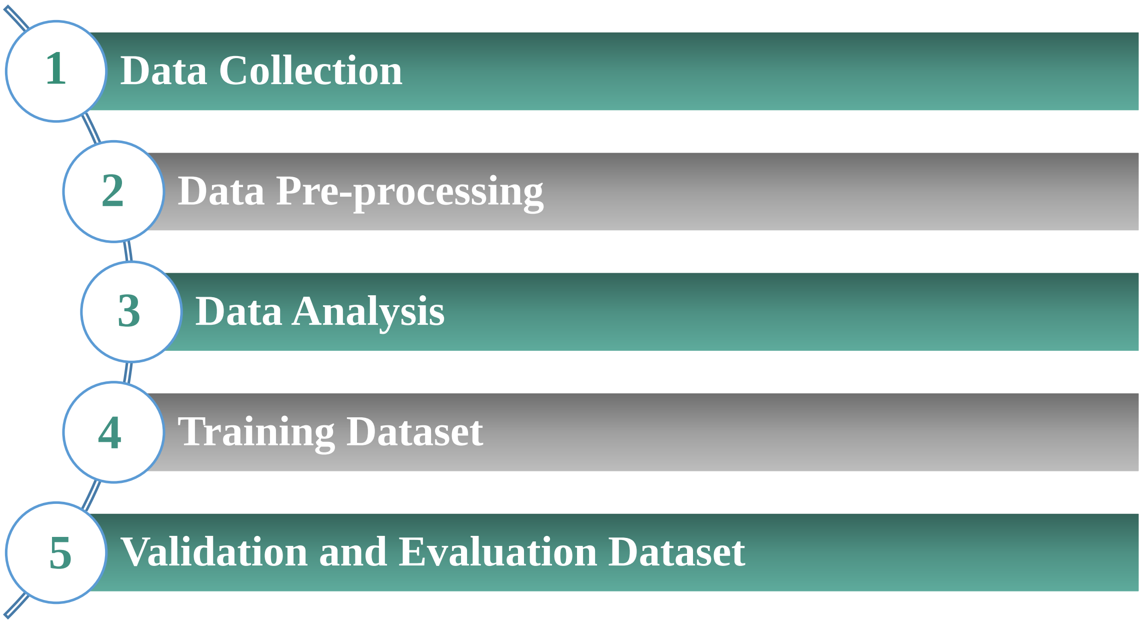


Figure 3.1: Steps of Proposed Methodology.

3.2.1 Data Collection

Arabic language is widely used in the world and is considered the most important between other languages because spoken by 422 million people in the world. The dataset collected in different fields for 8 subjects (Islamic, History, Geography, Biology, Computer, Geology, Chemistry, and Physics), each question has 3 typical answers, all essay questions in the Arabic language. Gathered a clear dataset from the original place, identify sources, select the dataset. The dataset collected from the teacher's book is published by the Jordan Ministry of Education's in 2019 for two-level (eleventh grades, and twelfth grades).

3.2.2 Data Structure

The dataset collected in different fields (science, geography, etc). Each question contains 3 typical answers. The dataset was converted from Microsoft Excel (.xlsx) to a comma Separated Values CSV file and ARFF file. Table 3.1 shows a data structure as follows:

Table 3.1: Dataset Structure.

Essay_id	Essay-question	Answer	Rate
1	ما المقصود بالبحث العلمي	هو أسلوب منظم في اكتشاف المعرفة والوصول إليها باستخدام الأدوات الموضوعية المتاحة التي لا تتأثر بذاتية الباحث ومشاعره.	5
1	ما المقصود بالبحث العلمي	هو أسلوب منظم في اكتشاف المعرفة.	2
1	ما المقصود بالبحث العلمي	فارغ.	0
2	ما المقصود بالحوار	هو أسلوب تواصل يمارسه الأفراد في أثناء تبادل الآراء والأفكار بهدف الوصول الى الصواب وفق ضوابط وأسس سليمة.	5
2	ما المقصود بالحوار	هو أسلوب تواصل يمارسه الأفراد في أثناء تبادل الآراء.	2
2	ما المقصود بالحوار	فارغ.	0

That dataset collected in Microsoft Excel after that to convert Comma Separated Values (CSV), it is a file used to store tables and it uses commas so that each line represents several columns and represents text or numbers.

3.2.3 Data Pre-processing and Analysis

In the Arabic essays, the pre-processing is more challenging, but it is very useful because it cleans up the dataset and remove any noisy and unnecessary words like remove stop words, remove conjunction, and so on.

some steps of the preprocessing, the first step tokenization is divided the large text too small pieces called (terms), the terms separated by punctuation like (Quotation marks, comma, and space), without attention what the meaning of these words, and the relation between them.

The second step of preprocessing was normalized dataset values in the given dataset use to eliminate the redundant useless word, the scale the dataset in the range [0,1]; where 1 is the largest value and the 0 is the smallest value (Sallam, Mousa, & Hussein,2016).

The third step of preprocessing was stop-word removal remove unusable and unusual dataset, remove special characters like (&,% , (), $ @, \,) remove conjunction like (so بالتالي , but لكن, for الى/عن/على), remove pronouns like (as we نحن, it هو/هي and you انت,), remove stop word like (at على/في, of عن, until حتى), and replace some characters like (إ ,آ ,أ, with ا, and ـه ,ـة, ة, ه with ه), this step it works to save the most important words.

The last step of preprocessing was stemming, which was the process of deleting the prefix and suffix from words it used to reduce all words retain the origin it's called (root), it is very important and useful. For example, stemming " " ," جماعي " ," جمعة " ," مجموعة" ," اجتماعي " ," جامعة " ," تجمع" ," اجتماع" ," اجتماعية" ," مُجمع" ," جُموع " ," استجمع" ," جماعة" ," مجاميع " ," اجمع", all these words the same root "جمع". The Arabic language has 11,347 roots, this study used stemmer valid in the WEKA tool is (Arabic light stemmer) to reduces the number of words (Al-Omari, & Abuata, 2014).

3.2.4 Training Dataset

At this stage, create the model to train the dataset in the machine learning by giving it a set of typical questions, answers, and grading. We train the model in machine learning on inputs and outputs. When the training process increases and inputs a large set of data, the accuracy gradually will increase, performance continues to increase until the dataset was ready for the testing process.

3.2.5 Dataset Validation and Evaluation

In the second part of the dataset, is the testing, the dataset was selected and randomly chosen in the testing process. the divided dataset in two parts, there was part of the dataset for training and another part for testing, and take the largest part of the training process, for example, 80% it is training, and selected 20% the testing process, where the inputs are available but the outputs are not available and they should be predicted the class during the application.

The testing process is only through evaluating the performance of the algorithm and selecting the best from many algorithms, and evaluation is done through who is the best accuracy. The following Figure 3.2 shows how the split dataset (Tarang, 2017) .

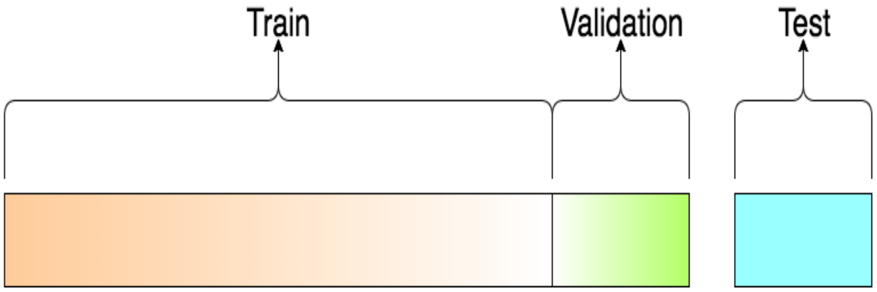


Figure 3.2: The splits of the Dataset (Tarang, 2017).

The dataset evaluated using machine learning and its effectiveness and accuracy evaluated using the following equations: (Davis, & Goadrich,2006).

- **Recall or True positive rate (TP Rate)**: The ability of a model to predict probabilities to each class (actual positive), it is calculated in equation 2:

$$\boxed{\textbf{Recall =TP/ TP+FN}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.. (2)$$

- **Precision**: Average rates for classification and predicted classifier is a correctly (predicted positive), calculated in follows equation 3:

$$\boxed{\textbf{Precision =TP/ TP+FP}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

- **False positive rate (FP Rate):** Average rates for a classification and predicted classifier is an incorrectly, it is calculated in equation 4:

$$\boxed{\textbf{FP Rate =FP/TN+FP}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

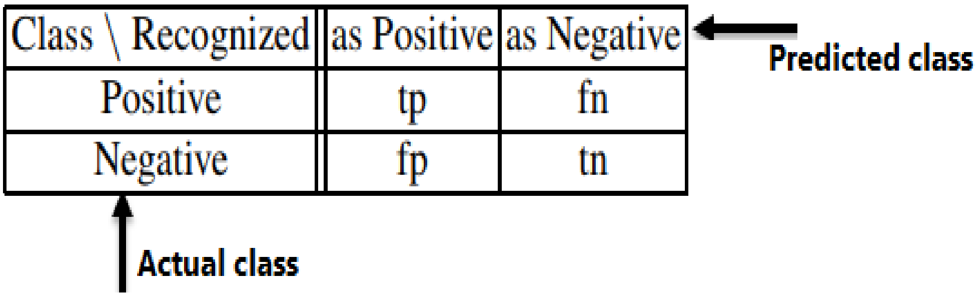
Where (TP): Actual positive and predicted positive, (FP): Actual negative and predicted positive, (FN): Actual positive and predicted negative, (TN): Actual negative and predicted negative.

- **Mean Absolute Error (MAE):** measures the average of the errors in the dataset for each algorithm (NB, J48, Meta) using equation 5.

$$\text{MAE} = \frac{1}{n}\sum_{i=0}^{n} | e_i | \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

Where n= The number of errors , ei, i = Real values (1, 2, ……., n).

- **Confusion Matrix**: It is a table that contains information about the actual classifications and predictive classifications predicted by the workbook. Each column in the matrix represents expected category and each row represents actual separation. Figure 3.3 shows arrays containing different numbers called Confusion Matrix (Bowes, Hall, & Gray,2012).



Class \ Recognized	as Positive	as Negative
Positive	tp	fn
Negative	fp	tn

Predicted class
Actual class

Figure 3.3: Confusion Matrix. (Sokolova, Japkowic, & Szpakowicz, 2006).

Chapter Four

Implementation and Evaluation Results

Chapter Four

Implementation and Evaluation Results

4.1 Introduction

In this chapter, we conducted a set of experiments on the collected dataset using the machine learning algorithms (Naive Bayes, decision tree, and meta classifier) to evaluate the validity of the dataset and measuring accuracy after that comparing results. The experiments were conducted using machine learning by using WEKA tools.

4.2 Collecting Dataset

In this thesis, there are a set of questions and typical answers from the teacher's book are published by the Jordan Ministry of Education's in 2019 for two-level (eleventh grades, and twelfth grades). The dataset collected in different fields for 8 subjects (Islamic, History, Geography, Biology, Computer, Geology, Chemistry, and Physics) and made spellcheck. The dataset structure as follows:

Essay_id: Number of the question.

Essay_qustion: The question is different in length as the number of words ranges (3-25 words).

Answer: The answer is different in length as the number of words ranges (0-100 words). Each topic is individually.

Rate: each question contains 3 typical answers; First, the set of answers that get full marks together (5/5). Second, set of answers that missing some important words that get part of the mark (2/5), last one the answer doesn’t get mark because of empty answer or rubbish data (0/5). Example of dataset for two topics, each one has a question, a group of answers, and rate of each answer. The structure as follows:

- **Subject: Islamic**

Q1: ما المقصود بالبحث العلمي

M1:هو أسلوب منظم في اكتشاف المعرفة والوصول إليها باستخدام الأدوات الموضوعية المتاحة التي لا تتأثر بذاتية الباحث ومشاعره

M2: هو أسلوب منظم في اكتشاف المعرفة

M3: فارغ

- **Subject: Geology**

Q1: ما العلاقة بين جتا زاوية سقوط الأشعة وتدفق الأشعة الساقطة (درجة الحرارة)

M1: العلاقة طردية ، فكلما زاد جتا سقوط الأشعة زاد تدفق الأشعة الشمسية الساقطة ، وبالتالي تزداد درجة الحرارة

M2:تزداد درجة الحرارة.

M3:فارغ

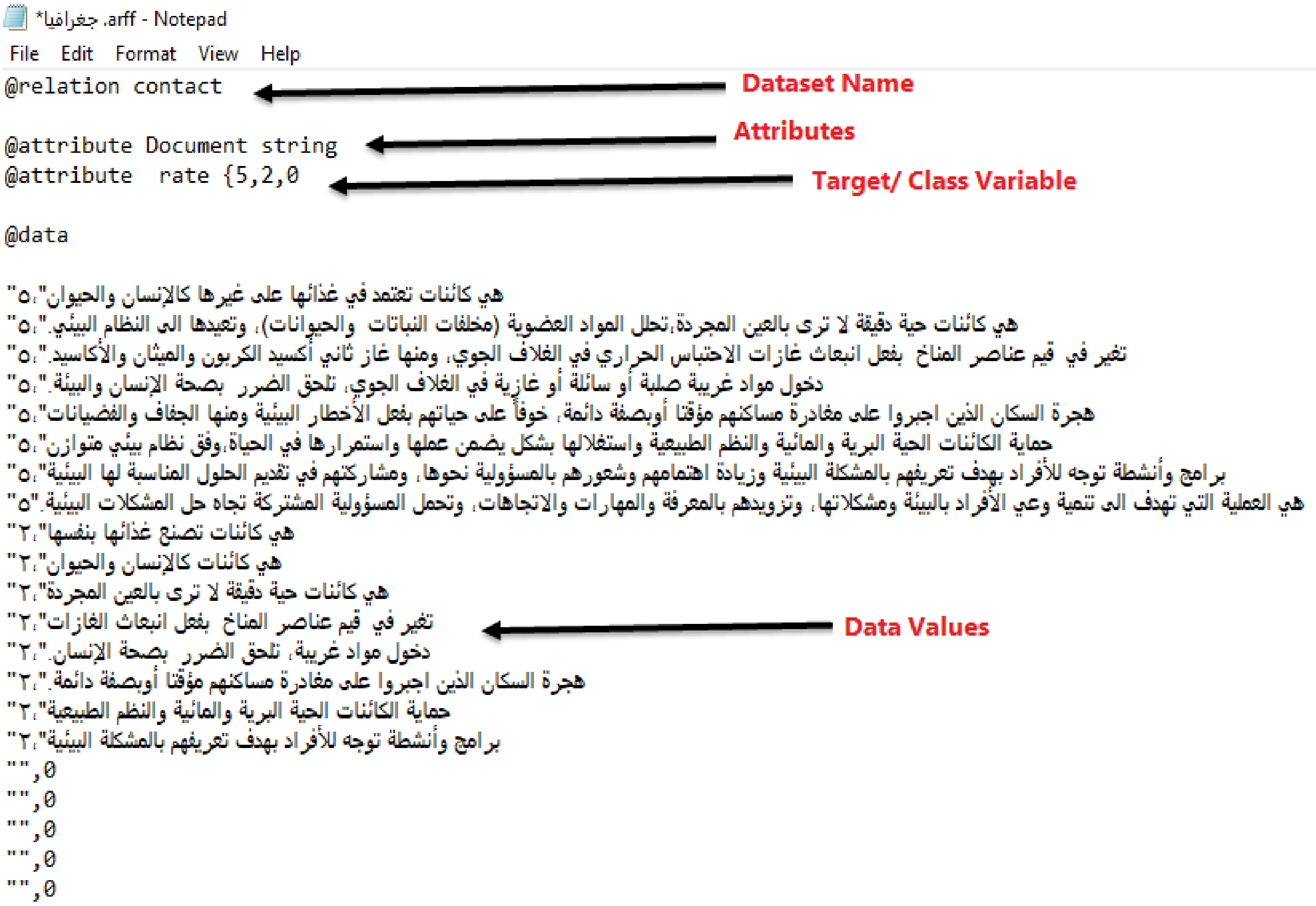
- **<u>Rate</u>**

M1: 5/5

M2: 2/5

M3: 0 /5

The data was converted to a format that can be easily used into machine learning like (CSV) and converted to (Attribute-Relation File Format ARFF file) because the used experimental tool is WEKA. The data is written is as follows: first step dataset name where each file was named a specific subject (Biology, Physics, Chemistry, Geology, Computer, Geography, History, Islamic). The second step was when the attributes of the dataset were entered as string. The third step, where target class variables were three to represent the grades of the essay as determined by human grader as follows (5,2,0). The last step wrote the dataset. Figure 4.1 shows a sample of the dataset for Certain topics.



```
@relation contact

@attribute Document string
@attribute  rate {5,2,0

@data

"هي كائنات تعتمد في غذائها على غيرها كالإنسان والحيوان"،٥
"هي كائنات حية دقيقة لا ترى بالعين المجردة،تحلل المواد العضوية (مخلفات النباتات  والحيوانات)، وتعيدها الى النظام البيئي."،٥
"تغير في  قيم عناصر المناخ  بفعل انبعاث غازات الاحتباس الحراري في الغلاف الجوي، ومنها غاز ثاني أكسيد الكربون والميثان والأكاسيد."،٥
"دخول مواد غريبة صلبة أو سائلة أو غازية في الغلاف الجوي، تلحق الضرر  بصحة الإنسان والبيئة."،٥
"هجرة السكان الذين اجبروا على مغادرة مساكنهم مؤقتا أوبصفة دائمة، خوفاً على حياتهم بفعل الأخطار البيئية ومنها الجفاف والفضيانات"،٥
"حماية الكائنات الحية البرية والمائية والنظم الطبيعية واستغلالها بشكل يضمن عملها واستمرارها في الحياة،وفق نظام بيئي متوازن"،٥
"برامج وأنشطة توجه للأفراد بهدف تعريفهم بالمشكلة البيئية وزيادة اهتمامهم وشعورهم بالمسؤولية نحوها، ومشاركتهم في تقديم الحلول المناسبة لها البيئية"،٥
"هي العملية التي تهدف الى تنمية وعي الأفراد بالبيئة ومشكلاتها، وتزويدهم بالمعرفة والمهارات والاتجاهات، وتحمل المسؤولية المشتركة تجاه حل المشكلات البيئية."٥
"هي كائنات تصنع غذائها بنفسها"،٢
"هي كائنات كالإنسان والحيوان"،٢
"هي كائنات حية دقيقة لا ترى بالعين المجردة"،٢
"تغير في  قيم عناصر المناخ  بفعل انبعاث الغازات"،٢
"دخول مواد غريبة، تلحق الضرر  بصحة الإنسان."،٢
"هجرة السكان الذين اجبروا على مغادرة مساكنهم مؤقتا أوبصفة دائمة."،٢
"حماية الكائنات الحية البرية والمائية والنظم الطبيعية"،٢
"برامج وأنشطة توجه للأفراد بهدف تعريفهم بالمشكلة البيئية"،٢
"",0
"",0
"",0
"",0
"",0
```

Figure 4.1: Sample of the Dataset

The dataset in this study contains 1003 questions and 3009 answers, the dataset collected in different fields for 8 subjects, there is distributed several questions for each topic as shown in table 4.1.

Table 4.1: Number of Questions

Topic	Number of questions	Number of answers
Islamic	322	966
History	190	570
Geography	181	543
Biology	80	240
Computer	88	264
Geology	81	243
Chemistry	35	105
Physics	26	78
Total	1003	3009

The Arabic language has 28 letters, and 22 have litters on different shapes like (ت، تــ ، ـتــ); therefore, the Arabic language contained a large number of words. Dataset analysis shows a variation in the number of words for more than one subject as the number of words is different for each subject. Also, remove the repeated word and keep the unique words were extracted from all topics. Table 4.2 shows the total number of words and the number of unique words for each subject, figure 4.2 shows the graph of the diversity of words. The collected dataset shows a diversity of words and show the number of words of each topic, can depict that each subject has different words and the number of words can be approximately 35000 and unique words about 10000.

Table 4.2: Number Unique Words

Topic	Number of words	Number of unique words
Islamic 12	7192	1050
Islamic 11	2241	774
Geography 12	4499	1307
Geography 11	3261	1118
History	3541	1229
Biology	3779	1361
Computer	4257	1578
Geology	4747	1140
Chemistry	1186	450
Physics	912	357

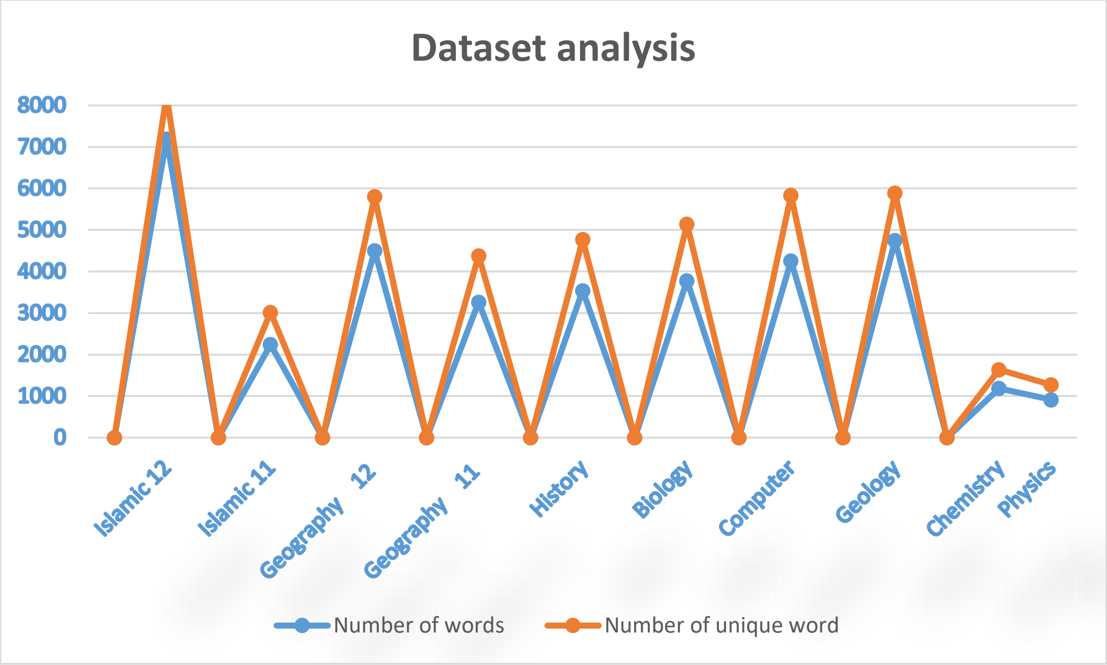


Figure 4.2: The Diversity of Words

Table 4.3, shows that there is a diversity when it comes to the number of sentences. We can depict that each answer has for on average two, but also some answers have three, to six sentences as illustrated in figure 4.3. Accordingly, for the 3009 answers of the dataset there is in total about 4338 sentences.

Table 4.3: Number of sentences

sentences	Number of sentences
1	3028
2	608
3	457
4	127
5	82
6	36

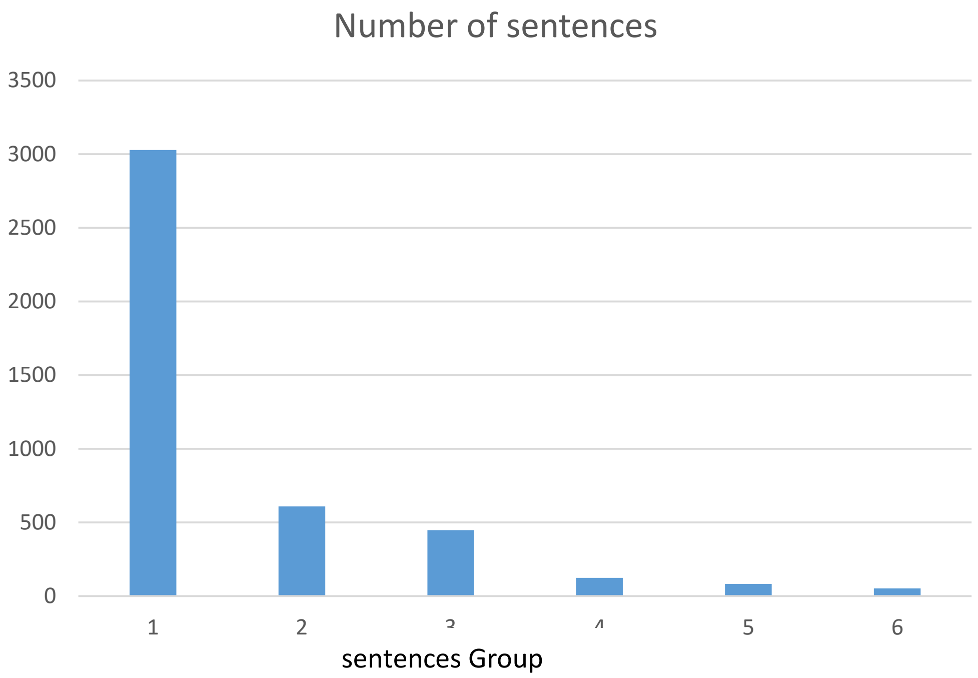


Figure 4.3: Number of Sentance.

Moreover, the mean of sentences for the whole dataset is 1.4431 as illustrated in figure 4.4,which calculated as in equation 6. In addition, the standard deviation = for the sentences is 0.8354, it is calculated as in equation 7. which statistically shows that the number of sentences of proposed dataset should be between 1 and 3 for each answer.

$$\text{mean} = \frac{\mathbf{S}}{\mathbf{N}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

Where S = The sum of the numbers answers, N = the number of sentence (1, 2, ………., n).

$$\text{Std}=\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\,x_i - \overline{x})^2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (7)$$

Std=Standard deviation, N=the number of sentences ,$\{x_1, x_2, \ldots \ldots \ldots . \ldots x_i\}$= the observed values of the sample items, $\bar{x}$ = the mean value of these observations.

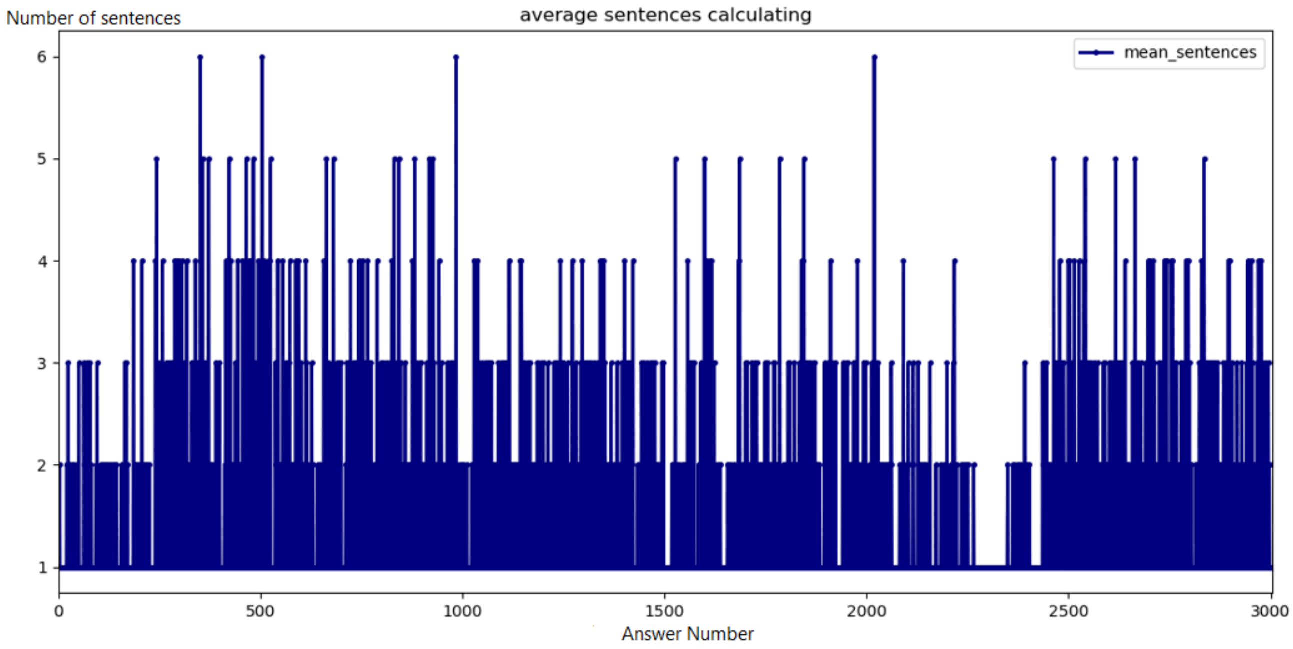


Figure 4.4: The mean of sentences in each answer.

4.3 Experimental Evaluation of The Benchmark

We have evaluated the dataset by using machine learning algorithms (Naïve Bayes, decision tree, meta), the experiment conducted using the WEKA tool (Waikato Environment for Knowledge Analysis), it is open-source software is written in Java and, where it can be applied easily and help in data analysis and data mining (Wahbeh & Al-Kabi,2012). The experiment is conducting as follows.

4.3.1 Pre-processing of the benchmark

Import dataset and pre-processing an important step where the dataset requires to delete unwanted words or delete the error words, its use to show the best results for the dataset. The dataset can be downloaded in ways by open file or open URL.

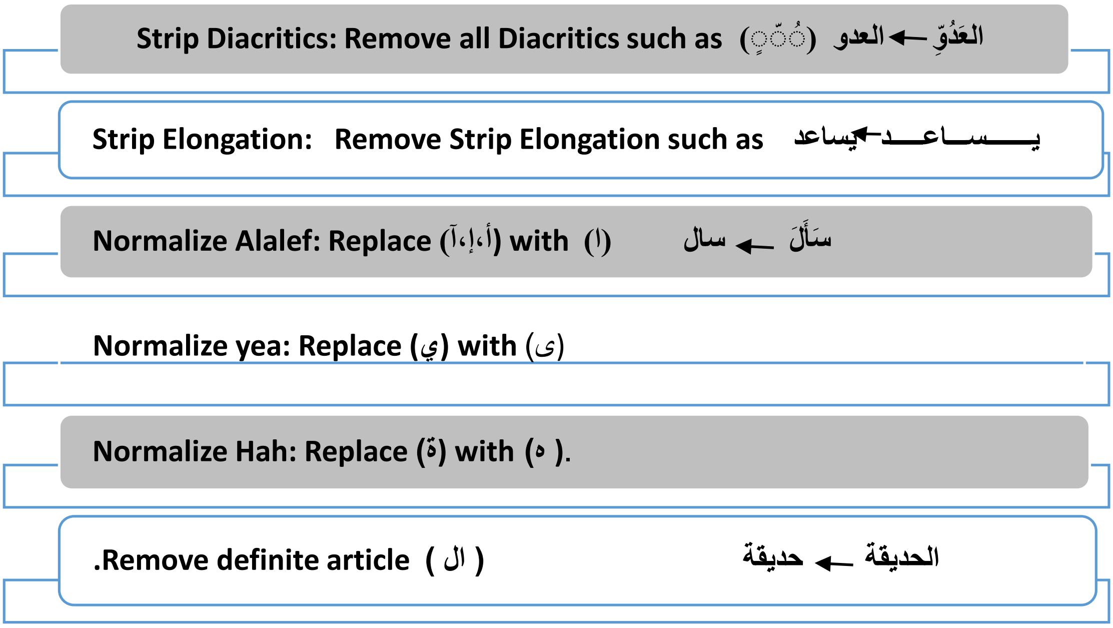
After loading the dataset feature selection by using the filter in the WEKA tools is important for data pre-processing, the filter converts dataset from format to another. WEKA filters are divided into two types: (Bouckaert, & other, 2010).

Supervised filters: are filters that dependable on data classification on class attributes. Unsupervised filters: are undependable on class attributes in their operations the best filter use for text classification, it is (String to Word Vector).

Use it to convert string attributes into a set of numeric attributes representing word occurrence information from the text contained in the strings. The steps of the preprocessing as follows.

4.3.1.1 Normalization

Normalization is important for processing Arabic text, the following figure 4.5 as shows examples of normalization Arabic text: (Sallam, Mousa, & Hussein,2016)



Strip Diacritics: Remove all Diacritics such as (ٍُّ) العَدُوِّ ← العدو

Strip Elongation: Remove Strip Elongation such as يـــــســـاعـــد ← يساعد

Normalize Alalef: Replace (أ،إ،آ) with (ا) سَأَلَ ← سال

Normalize yea: Replace (ي) with (ى)

Normalize Hah: Replace (ة) with (ه).

.Remove definite article (ال) الحديقة ← حديقة

Figure 4.5: Normalization Arabic Text

4.3.1.2 Stop words

The words that were repeated in texts such as (من، حيث، الى، ....) and it was desirable to ignore them to don't affect classification, used to reduce the time during classification. table 4.4 shows examples of Arabic stop words (Nassar, Taha, &Nazmy,2007)

Table 4.4: Arabic Stop Words

ءَ	أنت	أي	إذ	؟	اللذان	بهذا	عليه	مرّة
آ	انا	أيا	إذا	االا	اللذين	ة	عند	يوم
آل	أنتم	ؤ	إذما	ااالتي	امام	تم	عنها	وجد
الإ	أنتما	أف	إحدى	التي	امس	تلك	قال	وأبو
أم	أنتنّ	إمّا	إليك	الذي	اليوم	ثمّ	قبل	هلم
أما	أنتن	إن	إليكنّ	الذين	ايضاَ	حتى	كأن	هيهات
ان	أنه	إنَّ	اياك	اللتان	بعد	حبذا	كثيرا	هَؤلاء
أه	أو	إي	اياكم	اللتين	بل	حقا	لازال	هَاتِه

4.3.1.3 Stemming

Stemmer is important for processing text. Used stemmer to delete all of the prefixes in words and used to replace the words that have the same meaning and find words that are derived from the same stem/root (Al-Omari, & Abuata, 2014).

The Arabic language is more difficult than the Western language, and the Arabic language has 11,347 roots and it was difficult to find a proper stemmer to be used in this thesis. Thus, in this thesis used the popular stemmer used in most Arabic NLP research such as ISRI stemmer and the Arabic light stemmer in WEKA as presented by (AL-Ameed, et al., 2005), a sample of stemmed teams in the dataset as shown Table 4.5.

Table 4.5: Sample Stemming

Word	Postfix	Suffix	Root (Core)	Prefix	Antefix
الــســاعــة		ــة	ساع	ال	
لــيــحدثــونــكــم	كــم	ون	حدث	يـــ	لــ
بـــالــتــالــي		ي	تـــال	الــ	بــ
الـــفـطــريـــات	ات	ي	فطر	الــ	
العاطفـــة		ة	عطف	الـــ	
الطــلبـــات		ات	طلب	الــ	

4.3.2 Feature Extraction

Feature Extraction useful to dimensionality reduction, where used TFIDF it is one of the main techniques of information processing is used for the text feature weight. It is very important to show the importance of a term and improves the classification accuracy.

TF-IDF (term frequency-inverse document frequency) is a common method used to measure the frequency of the word in texts and documents, words that are frequently occurring get high frequency but the words that are not frequently occurring get low frequency calculated using following equation 8. (Bin, & Yuan ,2012).

$$\text{TFIDF} = \text{TF} * \text{IDF} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(8)}$$

Where (TF): Term frequency, (IDF): Inverse-Document-Frequency.

Term frequency (TF): measures how frequently a word in text, using following equation 9, **Inverse-Document-Frequency (IDF): measures general** importance the informativeness of term, it's called a relative weightage, using following equation 10.

$$\text{TF(i, j)} = \frac{Term\ i\ frequcy\ in\ document\ j}{Total\ words\ in\ document\ j} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(9)}$$

$$\begin{aligned} \text{idf} &= -\log p(t|D) \\ &= \log \frac{1}{p(t|D)} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(10)} \\ &= \log \frac{N}{|\{\ d\epsilon\ D\colon t\epsilon\ d\}\ |} \end{aligned}$$

where t denotes to the term**,** D: document, N: number of the total document set.

4.4 Experimental Result and Discussion

The experiments were conducted used WEKA tools, it has short-list for classification algorithms available as follows: Naive Bayes, Decision trees, meta classifier (Brownlee, 2016), and can use four testing options to choose any one of them:

- **Using Training set:** The classification takes a training dataset that was given label to the dataset.
- **Supplied test set:** The classification takes the same training dataset but does not have a label to the dataset and begins to guess the label for it.

- **Cross-validation:** Dividing the dataset into training and testing groups. For example, if using this method and the dataset is 100, then gives the value of (folds) 10, the dataset may produce a small group is divided into nine groups for training, and only the last group for the test.
- **Percentage Split:** The dataset is divided into two groups, a training group, and a test group. This study used to classification more than one algorithm for the dataset when chooses dataset to training must choose the same dataset without a label in the testing.

4.4.1 Naive Bayes classifier

The first classifier used in this thesis to evaluate and test the dataset is the NB. The results obtained show that the percentage of essays that were graded correctly is 79.0483% compared to the 20.9517% of essays that were graded incorrectly. Moreover, as shown in table 4.6, metrics were also used to verify the performance of NB on our dataset for each class of grades. Figure 4.6 shows the graph of the results of the ROC area for NB where the ROC is above the threshold which means that NB performance was accurate for grading the essays.

Table 4.6: Evaluation Metrics Results for Naive Bayes.

	Evaluation Metrics				
	Precision	**Recall**	**F-Measure**	**ROC Area**	**Class**
NB	0.869	0.785	0.825	0.909	5
	0.795	0.810	0.802	0.882	2
	0.922	1.000	0.959	0.984	0

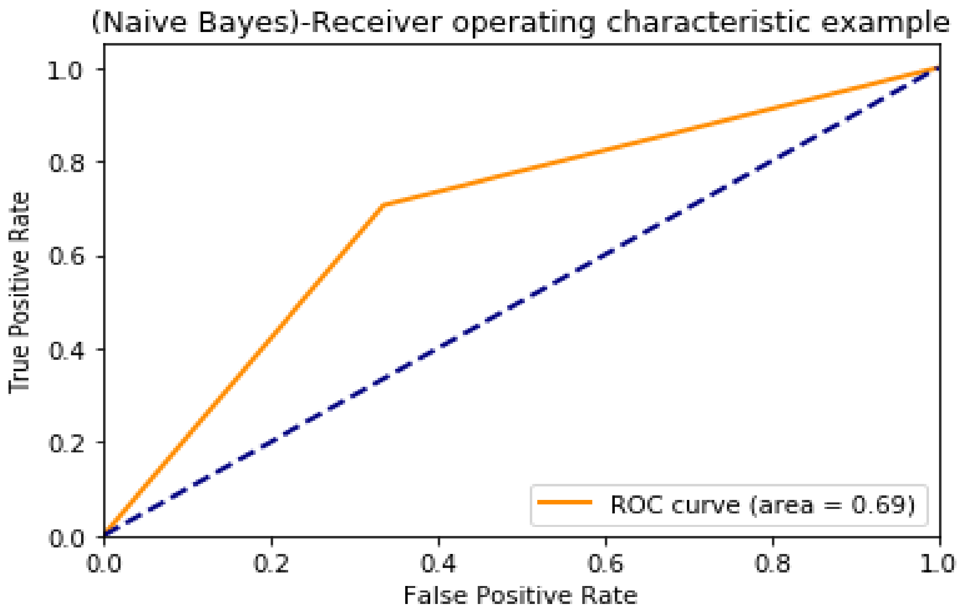


Figure 4.6: The Results of the ROC Area for NB.

4.4.2 J48 classifier

The second classifier used in this thesis to evaluate and test the dataset is the J48. The results obtained shown that the percentage of essays that were graded correctly is 81.2855% compared to the 18.7145% of essays that were graded incorrectly. Moreover, as shown in table 4.7, metrics were also used verify the performance of J48 on our dataset for each class of grades. Figure 4.7 shows the graph of the results of the ROC area for J48, where the ROC is above the threshold which means that J48 performance was accurate for grading the essays.

Table 4.7: Evaluation metrics results for Decision tree.

	Evaluation Metrics				
	Precision	**Recall**	**F-Measure**	**ROC Area**	**Class**
	0.849	0.755	0.799	0.895	5
J48	0.759	0.634	0.691	0.802	2
	0.769	1.000	0.869	0.930	0

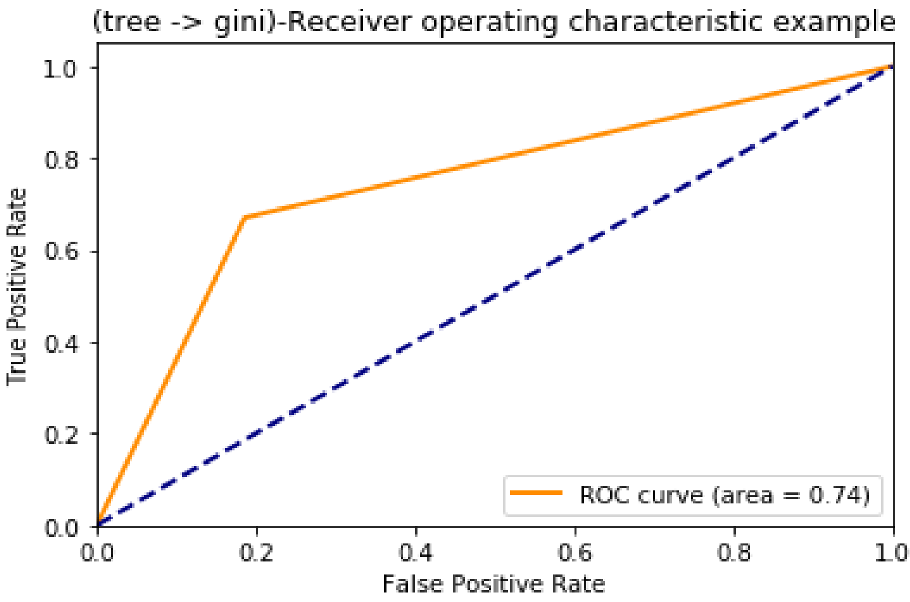


Figure 4.7: The Results of the ROC Area for J48.

4.4.3 Meta classifier

the third classifier used in this thesis to evaluate and test the dataset is the Meta classifier. The results obtained shown that the percentage of essays that were graded correctly is 86.1151% compared to the 13.8849% of essays that were graded incorrectly. Moreover, as shown in table 4.8, metrics were also used to verify the performance of the meta classifier on our dataset for each class of grades. Figure 4.8 shows the graph of the results of the ROC area for meta, where the ROC is above the threshold which means that Meta classifier performance was accurate for grading the essays.

Table 4.8: Evaluation metrics results for Meta classifier.

	Evaluation Metrics				
	Precision	**Recall**	**F-Measure**	**ROC Area**	**Class**
Meta	0.754	0.764	0.759	0.860	5
	0.747	0.690	0.718	0.746	2
	0.935	1.000	0.967	0.983	0

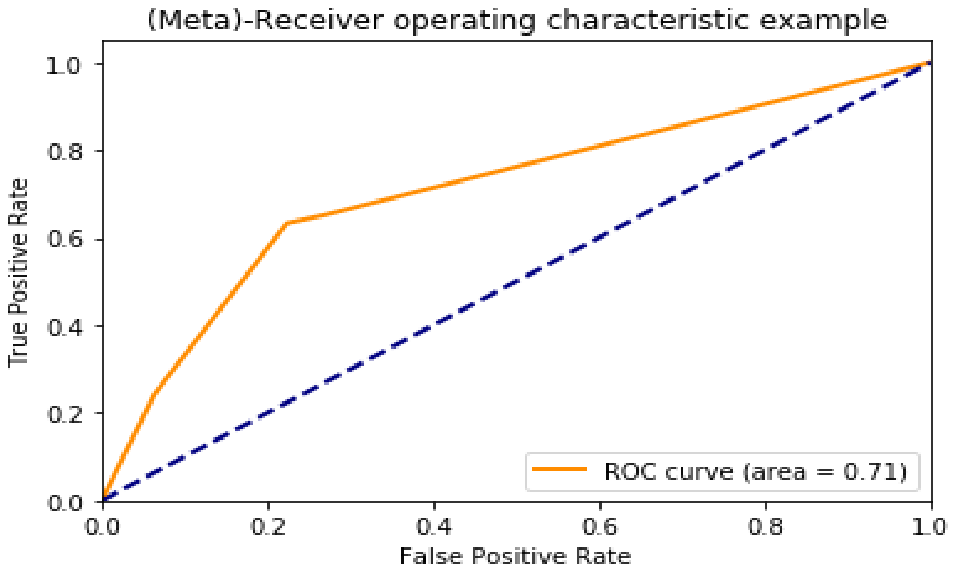


Figure 4.8: The Results of the ROC Area for meta.

Table 4.9 summarize the accuracy of the classifiers (J48, NB, Meta), while Figure 4.9 shows the graph of evaluation results of machine learning algorithms (NB, J48, Meta), for the classification on a dataset. Accordingly, the best classifier in terms of accuracy was Meta classifier based on accuracy and Mean Absolute Errors (MEA) on the collected dataset in this thesis.

Table 4.9: Classifiers Accuracy.

	Accuracy		
Algorithms	**Correctly Classified Instances**	**Incorrectly Classified Instances**	**Mean Absolute Error**
NB	**79.0483%**	**20.9517%**	**0.1927**
J48	**81.2855%**	**18.7145%**	**0.141**
Meta	**86.1151%**	**13.8849%**	**0.115**

Mean Absolute Error (MAE): After using the machine learning algorithms (NB, J48, Meta) the result is (0.1927, 0.141, 0.115).

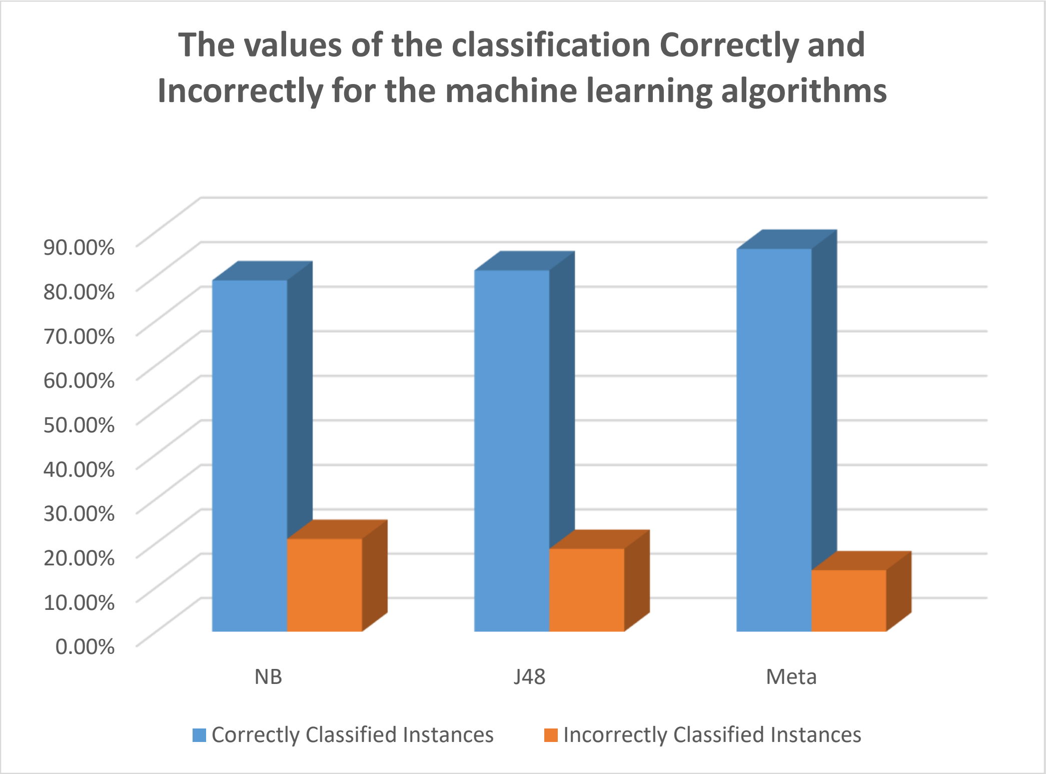


Figure 4.9: The Evaluation Results of ML Algorithms.

The best classifier in terms of accuracy was the Meta classifier. The results obtained shown that the percentage of essays that were graded correctly is 86%.

Chapter Five

Conclusion and Future Work

Chapter Five

Conclusion and Future Work

5.1 Conclusion

The Automated Essay Grading system is used to grade the students' essays and papers automatically without any bias or the interference of the teacher. This is a very significant tool for universities, colleges, and even schools. But, these types of AEGs are mostly developed for the English language and not for the Arabic natives. The simplest and foremost reason for this less development is the lack of availability of the Arabic dataset.

This thesis provides an Arabic language dataset that can be used for the Arabic AEGs by using the machine learning algorithms. The processed dataset is a collection of sets of questions and answers. These Q&A's are kept from the Teachers book and published by the Jordan Ministry of Education's in 2019 for two-level (eleventh grades, and twelfth grades). The dataset ranges from Islamic, History, Geography, Biology, Computer, Geology, Chemistry, and Physics and includes the 1003 questions and thrice the number of answers i.e. 3009.

The Arabic dataset is trained and tested on various machine learning models such as Decision Tree, Naïve Bayes, and Meta Classifier. The experimental results show the improved and reliable results predicted by the given models i.e. 81%, 79%, and 86% respectively.

5.2 Future Work

In the field of research, there are no limits or bars on the topic to end it on a benchmark. Whereas this is the beginning of new research based on the output of the present scenario but still, there are some recommendations and suggestions for future research such as:

1. Updating the dataset with more unique Arabic vocabularies.

2. The established dataset can be tested and evaluated using Deep learning methods.

References

Al Awaida, S. A., Al-Shargabi, B., Al-Rousan, T., Conroy-Beam, D., Buss, D. M.,Asao, K., ... & Abuarqoub, A. (2019). AUTOMATED ARABIC ESSAY GRADING SYSTEM BASED ON FScore AND ARABIC WORDNET. *Jordanian Journal of Computers and Information Technology (JJCIT)*, *5*(03).

Ali, M. ,., & Mohd, J. ,. (2013). Automatic essay grading system for short answers in English language. *Journal of Computer Science (JCS)* , 1369.

Al-Jouie, M. F., & Azmi, A. M. (2017). Automated evaluation of school children essays in Arabic. *Procedia Computer Science, 117*, 19-22.

Al-Omari, A., & Abuata, B. (2014). Arabic light stemmer (ARS). *Journal of Engineering Science and Technology*, *9*(6), 702-717.

Alsaleem, S. (2011). Automated Arabic Text Categorization Using SVM and NB. *Int. Arab J. e-Technol.*, *2*(2), 124-128.

Ayodele, T. O. (2010). Types of machine learning algorithms. *New advances in machine learning*, 19-48.

Aileen, S. (2018). Predictive Sales Analytics, a Project.

Batten, E. (1994). Computer Grading of Student Prose, Using Modern Concepts and Software. *JSTOR*, 127.

Bin, L., & Yuan, G. (2012, August). Improvement of TF-IDF algorithm based on Hadoop framework. In *Proceedings of the 2012 International Conference on Computer Application and System Modeling*. Atlantis Press.

Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., &Scuse, D. (2010). Weka manual for version 3-7-3. *The university of WAIKATO*, 327.

Bowes, D., Hall, T., & Gray, D. (2012, September). Comparing the performance of fault prediction models which report multiple performance measures: recomputing the confusion matrix. In *Proceedings of the 8th international conference on predictive models in software engineering* (pp. 109-118).

Brownlee, J. (2016). How to Use Classification Machine Learning Algorithms in WEKA. *Machine Learning Mastery*.

Chen, Z., & Liu, B. (2018). Lifelong machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *12*(3), 1-207.

Davis, J., & Goadrich, M. (2006, June). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning* (pp. 233-240).

.Fei, D., Yue, Z., & Jie, Y. (2017). Attention-based Recurrent Convolutional Neural. 153-157.

Gaurav, G. (2018). *Classification Algorithms in Machine Learning….*

Gomaa, W. H., & Fahmy, A. A. (2014). Automatic scoring for answers to Arabic test questions. *Computer Speech & Language*, *28*(4), 833-857.

Jakhar, D., & Kaur, I. (2020). Artificial intelligence, machine learning and deep learning:

definitions and differences. *Clinical and experimental dermatology, 45*(1), 131-132.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*, 3-24.

Larkey, L. S. (1998, August). Automatic essay grading using text categorization techniques. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 90-95).

Larry, R., Pepper, S., Andrew, P., Jennifer, B., Eric, B., Candac, e. C., . . . Edgar, B. (2012). *Handbook on Applying Environmental Benchmarking in Freight Transportation.*

Liu, B., Blasch, E., Chen, Y., Shen, D., & Chen, G. (2013, October). Scalable sentiment classification for big data analysis using naive bayes classifier. In *2013 IEEE international conference on big data* (pp. 99-104). IEEE.

Liu, J., Xu, Y., & Zhao, L. (2019). Automated essay scoring based on two-stage learning. *arXiv preprint arXiv:1901.07744*.K, D. (2019). *Top 5 advantages and disadvantages of Decision Tree Algorithm.*

Mathias, S., & Bhattacharyya, P. (2018, May). ASAP++: Enriching the ASAP Automated Essay Grading Dataset with Essay Attribute Scores. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

Matthew, M. (2018). *Frameworks for Approaching the Machine Learning Process.* KDnuggets News .

Mezher, R., & Omar, N. (2016). A hybrid method of syntactic feature and latent semantic analysis for automatic arabic essay scoring. *Journal of Applied Sciences*, *16*(5), 209.

Naresh, K. (2019). *Advantages and Disadvantages of Naive Bayes in Machine Learning.*

Nassar, H., Taha, A., Nazmy, T., & Nagaty, K. (2007, March). Classification of video scenes using Arabic Closed-Caption. In *Proceedings of the Third International Conference on Intelligent Computing and Information Systems, Cairo, Egypt*.

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. *IEEE Access*, *7*, 19143-19165.

Gaurav, G. (2018). *Classification Algorithms in Machine Learning….*

Nguyen, T. T., Luong, A. V., Van Nguyen, T. M., Ha, T. S., Liew, A. W. C., & McCall, J. (2019, July). Simultaneous meta-data and meta-classifier selection in multiple classifier system. In *Proceedings of the Genetic and Evolutionary Computation Conference* (pp. 39-46).

Niklas, D. (2018). *The Logistic Regression Algorithm.*

Rahul, S. (2017). *HOW THE NAIVE BAYES CLASSIFIER WORKS IN MACHINE LEARNING.* Data Science, Machine Learning.

Rudner, Lawrence., & Gagne, Phill. (2000). An overview of three approaches to scoring written essays by computer. *Practical Assessment, Research, and Evaluation*, *7*(1), 26.

Saad, M. K. (2010). The Impact of Text Preprocessing and Term. *The Islamic University - Gaza* .

Sallam, R. M., Mousa, H. M., & Hussein, M. (2016). Improving Arabic text categorization using normalization and stemming techniques. *Int. J. Comput Appl*, *135*(2), 38-43.

Semire, D. (2006). Automated Essay Scoring . *Turkish Online Journal of Distance Education-TOJDE* , 49.

Shehab, A., Faroun, M., & Rashad, M. (2018). An automatic Arabic essay grading system based on text similarity Algorithms. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, *9*(3), 263-268.

Sinka, M. P., & Corne, D. W. (2002). A large benchmark dataset for web document clustering. *Soft computing systems: design, management and applications*, *87,* 881-890.

Society, I. (April 2017). *Artificial Intelligence and Machine Learning: Policy Paper.*

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006, December). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence* (pp. 1015-1021). Springer, Berlin, Heidelberg.

Tarang, S. (2017). *About Train, Validation and Test Sets in Machine Learning.* Towards Data Science. (n.d.). *Training and Test Sets: Splitting Data.* Machine Learning .

Tastle, W., & Terzi, S. Meta Net (January 2013): A New Meta-Classifier Family. DOI: 10.1007/978-1-4614-4223-3_5

Wahbeh, A. H., & Al-Kabi, M. (2012). Comparative assessment of the performance of three WEKA text classifiers applied to arabic text. Abhath Al-Yarmouk: Basic Sci. & Eng, 21(1), 15-28

Witten, I. H., Frank, E., & Hall, M. A. (2005). Practical machine learning tools and techniques. *Morgan Kaufmann*, 578.