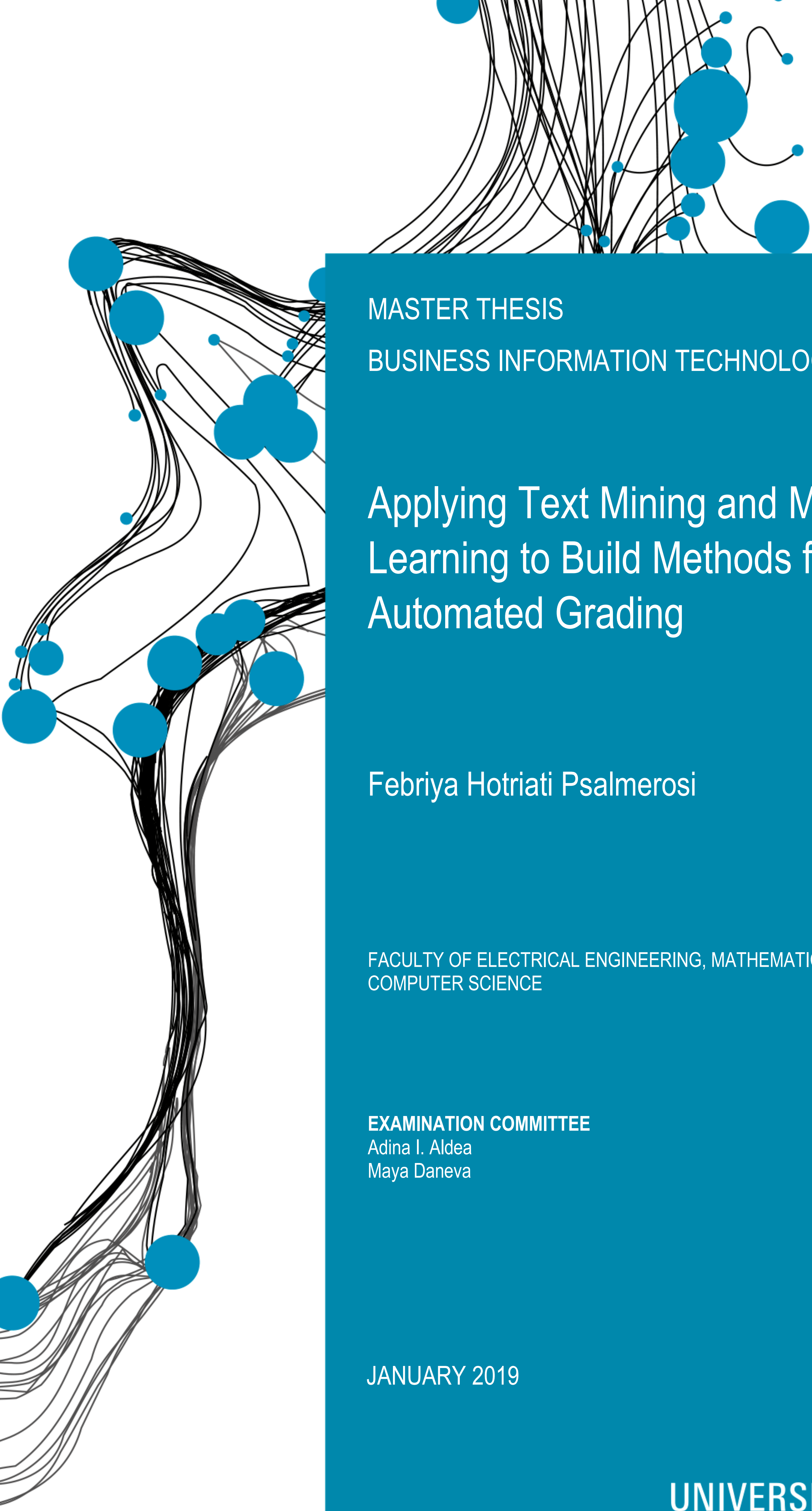


MASTER THESIS

BUSINESS INFORMATION TECHNOLOGY

Applying Text Mining and Machine Learning to Build Methods for Automated Grading

Febriya Hotriati Psalmerosi

FACULTY OF ELECTRICAL ENGINEERING, MATHEMATICS, AND COMPUTER SCIENCE

EXAMINATION COMMITTEE
Adina I. Aldea
Maya Daneva

JANUARY 2019

UNIVERSITY OF TWENTE.

ACKNOWLEDGEMENT

This thesis marks the end of my master study in Business Information Technology (IT) master program at the University of Twente. I have gained a lot of knowledge during the past two years, and I believe that what I have learned will enhance my career afterward. I also realize that I cannot achieve this point without the support of the people around me.

I would like to express my greatest gratitude to God for His providence and guidance during my study in The Netherlands. I also would like to say thank you to my country, especially to Ministry of Communication and Information (MCIT) of the Republic of Indonesia, that have funded my study. It is an honor to be one of MCIT scholarship awardee, and I will try my best to give my contributions to Indonesia later on. Moreover, I would like to thank my supervisors, Adina I. Aldea and Maya Daneva, for their guidance during my thesis project. Without your support and feedbacks, I cannot complete the project on schedule.

In this occasion, I also would like to say thank you to my family: my mom, Kak Boya, Bang Andre, Uti, and Kak Nita. Without your constant support and prayers, I cannot make it until here. I cannot thank you enough for pushing me to give the best of me in everything I do. I would like to express my sincere gratitude also for my best friends: Dhila, Tata, and Dona. Although we are far apart, thank you for always making time for me whenever I need a friend to talk.

Furthermore, I would like to show appreciation to my friends here, who make my days less lonely while living in this country. To my close friends: Fania, Fitri, Eva, Dzul, Victor, Kak Riris, and Ika. Thank you for all of the precious moments and laughs that we have shared. To my international friends, especially the ICF and choir friends: Agathi, Marilena, Somto, Bai, Dominik, Steven, Max, Lijun, Clement, Jan Maarten, Adwoa, Paul, Dink, Joshua, Miriam, and Jack. Thank you for making my Friday nights merrier, keeping me company during Christmases, and sharing your cultural experience with me. To my Indonesian friends: Bima, Yosia, Widi, Helena, Adrian, Bang Harry, Linda, Amanda, Erica, Hendry, Kevin, and other Indonesian fellows in Enschede, thank you for the friendship and delicious foods that can reduce the feeling of homesick.

And to other people that I cannot mention one by one, thank you for being a part of my journey during my study. I wish you all the best, and I hope we will meet again in the future.

Enschede, 28 January 2019

Febriya Hotriati Psalmerosi

ABSTRACT

Nowadays, machine learning and text mining have become an interesting topic for both research and practice. The impact of machine learning and text mining technologies is significant in any area of the business or the public sector, including education. Specifically, in education, one of the most interesting applications of these technologies is in the evaluation of students' tests' results that come out of open-question-based examination processes. This thesis responds to the trend to employ machine learning and text mining techniques in evaluation of students' responses to open questions. The present research is focused on the identification of the best approach to automate the grading of students' answers in open-question-based examination. To this end, we conducted a comparative study of a set of alternative methods.

In open-question-based examination, there are several types of open questions, however previous researches have been done for the essay and short answer only. This study explores the grading process as supported by machine learning and text mining techniques, regarding two types of open questions: (1) "mention and explain a couple of examples for different categories," and (2) "give a concise and valid argument about a given statement." Additionally, the present study focuses on finding better approaches for the small dataset (less than 50) in contrast to previously published literature which tends to investigate their method in a big dataset.

Therefore, current research provides several contributions to the theory. This study examines other open question types that have not been explored in previous works. This research also proposes techniques for automated grading system using combinations of text mining and machine learning for an automated grading system for the small dataset. Then, this study demonstrates the use of RapidMiner for automated grading implementation.

This study uses text mining and machine learning techniques to assess each question type. Unlike related works in this area, the present study does not aim to give a score to a student's answer, but to examine those characteristics of an open question that can be advantageous for automated grading. Therefore, this research provides several suggestions for lecturers about how to create a question that can be easily graded by an automated system and to determine the performance of the implemented technique for two types of question.

Current research evaluates the proposed method in two ways: (1) by doing an experiment, and (2) by conducting an evaluation survey from three lecturers in the University of Twente. The first type of open questions is examined by counting the number of examples mentioned in the answer and by employing a classification technique using Support Vector Machine. The related experiment findings show the acceptable result to identify the number of examples within a category, with the accuracy of more than 70%. Moreover, the produced classifier model identifies the examples to its category with the accuracy of more than 85% and correlation value more than 0.700. These values signify high likelihood that students' answers are similar and related to each other.

For the second type, this study implements sentiment analysis and clustering with X-Means algorithm. The Davies-Bouldin (DB) index and Silhouette index are applied to measure the performance of the clustering. The optimal number of clusters is 7, using Manhattan Distance with DB index, which is 0.334, and Silhouette index which is 1.332. Our analysis found that answer length is the most dominant factor in determining the clusters, and Term Document Matrix influences the results of the clustering.

This master thesis project used RapidMiner for the purpose of experimentation. All answer files are written in text files.

In addition to the experiment results, an evaluation survey based on the Unified Theory of Acceptance and Use of Technology (UTAUT) by Venkatesh et al. (2003) was performed. From the evaluation results, performance expectancy becomes the strong determinant of the behavioral intention to use the proposed method. The most negative feedback is self-efficacy construct as there is a possibility that all participants think introduction session is important before using the proposed method.

TABLE OF CONTENTS

ACKNOWLEDGEMENT..i

ABSTRACT ii

TABLE OF CONTENTS iv

LIST OF FIGURES...vi

LIST OF TABLES...vii

CHAPTER 1 – Introduction...1

1.1. Motivation 1

1.2. Problem Definition...2

1.3. Research Question...3

1.4. Research Methodology 3

1.5. Contributions...4

1.5.1. Contributions to Theory...5

1.5.2. Contributions to Practice 5

CHAPTER 2 – Literature Review...6

2.1. Open Questions 7

2.2. Machine Learning 8

2.3. Text Mining...9

2.4. Automated Grading 11

2.5. Datasets in Previous Work of Automated Grading 22

2.6. The relationship between Machine Learning, Text Mining, and Automated Grading................23

2.7. Tools Selection...23

CHAPTER 3 – Method..25

3.1. Mention a number of examples and explain 25

3.2. Ask opinion from the students..28

CHAPTER 4 – Experiment Design and the Implementation in RapidMiner 31

4.1. Dataset 31

4.2. Whole Process 32

4.3. Mention a number of examples and explain 34

4.3.1. Split the category and count the number of examples for each category 34

4.3.2. Check the answer content 37

4.4. Ask opinion from the students 40
4.4.1. Sentiment Analysis 40
4.4.2. Clustering 41
CHAPTER 5 – Results 45
5.1. Experiment Results 45
5.1.1. Mention a number of examples and explain 45
5.1.2. Ask opinion from the students 51
5.2. Evaluation Result 55
5.2.1. Performance Expectancy 58
5.2.2. Effort Expectancy 59
5.2.3. Facilitating Conditions 59
5.2.4. Self-efficacy 60
5.2.5. Attitude toward using technology 61
5.2.6. Behavioral intention to use 61
CHAPTER 6 – Conclusions 63
6.1. Main Finding 63
6.2. Discussion 64
6.3. Limitations 65
6.4. Future Work & Recommendation 65
REFERENCES 67
Appendix A – Figures of Process 72
Appendix B – Table of Result 76
Appendix C – Questionnaire design 84

LIST OF FIGURES

Figure 1 DSRM Process Model by Peffers 4
Figure 2 SLR Process 7
Figure 3 Proposed Method 25
Figure 4 First Process: Check the Amount of Examples & Category 26
Figure 5 Second Process: Check The Answer Content 26
Figure 6 Second Process: Inside Prepare Dataset Process 27
Figure 7 Second Process: Inside Create Word Vectors Process 27
Figure 8 Ranking Process 27
Figure 9 Ranking Process: Inside Calculate total_confidence Loop 28
Figure 10 Main Process of Question Type 2 28
Figure 11 Inside Create Word Vectors Process 29
Figure 12 Whole Process 33
Figure 13 Whole Process: Inside Process The Answers Loop 33
Figure 14 Whole Implementation of First Process 34
Figure 15 An Example of Loop Files Operator Results 35
Figure 16 Process Inside Loop Files Operator (1) 35
Figure 17 Process Inside Loop Files Operator (2) 36
Figure 18 Process of Counting Number of Categories and Examples 36
Figure 19 The Results of Answers Grouping 37
Figure 20 The Summary of Answers Grouping Results 37
Figure 21 Check The Answer Content: Training Process 38
Figure 22 Training Process: Specify the Directory in Process Document from Files Operator 38
Figure 23 Training Process: Inside Process Documents From Files Operator 38
Figure 24 Training Process: Answers Pre-processing Inside Cut Document (3) Operator 39
Figure 25 Training Process: Subprocess Inside Cross Validation Operator 39
Figure 26 Check The Answer Content: Evaluation Process 40
Figure 27 Sentiment Analysis Process 41
Figure 28 Process Inside Sentiment Analysis Operator 41
Figure 29 The Clustering Process 42
Figure 30 Answers Pre-processing Inside Process Documents from Data Operator 42
Figure 31 Scree plot of SVD Results 43
Figure 32 Clustering Implementation 43
Figure 33 Process Inside X-Means (2) Operator 44
Figure 34 Sentiment Analysis Results using AYLIEN Text Analytics 51
Figure 35 Polarity Distribution of Each Cluster 53
Figure 36 Maximum and Minimum Answer Length of Each Cluster 53
Figure 37 UTAUT Research Model (Venkatesh et al. 2013) 55
Figure 38 Descriptive Statistics of The Survey 58

LIST OF TABLES

Table 1 Previous Works in Automated Grading System 12
Table 2 The Summary of Previous Works based on Information Extraction Methods 15
Table 3 The Summary of Previous Works based on Machine Learning Methods 19
Table 4 Datasets Used in Automated Grading System 22
Table 5 Comparison of Tools Used in Automated Grading System 23
Table 6 Accuracy Results of Counting Number of Examples and Categories 46
Table 7 Threshold Value of Measurement 48
Table 8 Results of Content Checking 49
Table 9 Confusion Matrix of Training Set Second Question 49
Table 10 Confusion Matrix of Testing Set Second Question 49
Table 11 Top 5 of DB and Silhouette Index Results 51
Table 12 Silhouette Index for 7 and 10 Clusters 52
Table 13 Summary of Clustering Results of 7 Clusters with Manhattan Distance 52
Table 14 Clustering Results Compared to Human Score 52
Table 15 Constructs Summary in Estimating UTAUT 55
Table 16 Descriptive Statistics of The Survey Results 57
Table 17 Performance Expectancy Survey Results 59
Table 18 Effort Expectancy Survey Results 59
Table 19 Facilitating Conditions Survey Results 60
Table 20 Self-efficacy Survey Results 60
Table 21 Attitude Toward Using Technology Survey Results 61
Table 22 Behavioral Intention to Use Survey Results 61

CHAPTER 1 – INTRODUCTION

This chapter discusses the motivation behind the research, the problem definition, the research questions, the research methodology, and the contribution of the research.

1.1. Motivation

Multiple-choice and open questions are the most popular exams used in higher education to measure student's understanding during the learning process (Ozuru, Briner, Kurby, & McNamara, 2013; Stanger-Hall, 2012). Multiple-choice questions basically are built from a question and several alternative responses, which contain single or multiple correct answer(s) (Swartz, 2010). On the other hand, open questions elicit students to construct their own answers in a couple of sentences or paragraphs (Swartz, 2010; Wolska, Horbach, & Palmer, 2014). The examples of multiple-choice questions are true-false statements, matching, and traditional type (select correct answers from offered options), while open question exams include essays, long or short answers, case study, and reports (Swartz, 2010).

There are reasons why one type is preferred than the other. Lecturers use multiple-choice than open questions as a final assessment because it is easy to score, provides fast grading in large classes, and can fit more questions (Stanger-Hall, 2012). An experiment by Funk and Dickson (2011) revealed that students performed better in multiple-choice than open question test, but the performance result might overestimate students' understanding level of the course. Students might get the correct answer by guessing or unintentional hints in the alternative responses and not because of the students' competency (Funk & Dickson, 2011; Stanger-Hall, 2012; H. C. Wang, Chang, & Li, 2008).

Science education should engage students' abilities in independent thinking, problem-solving, planning, decision-making, and group discussion (H. C. Wang et al., 2008). Multiple-choice tends to have difficulty in examining students critical-thinking skills than open questions (Funk & Dickson, 2011) because they just have to select the correct answer from the alternatives and do not need to construct the answer in their own thought. Using open questions can help teachers to distinguish the level of understanding for each student from the quality of the answer.

Moreover, in open question exams, students are encouraged to prepare thoroughly and study more efficiently (Pinckard, McMahan, Prihoda, Littlefield, & Jones, 2009) because they are expected to answer in depth of knowledge and a wider range of thinking (Stanger-Hall, 2012). The open question reveals students' ability to integrate, synthesize, design, and communicate their thought (Roy, Narahari, & Deshmukh, 2015). The teachers can observe whether the students achieve the objective of the course or not by inspecting at how the students are applying their concepts and comprehension into a real problem.

Consequently, open question assessments have more values than multiple choice in measuring student comprehension of a problem. However, marking manually open questions exam requires a lot of resource in the matter of time. Grading an open question assessment need a lot of time because the teacher has to read each answer carefully. Each student might have a different way to answer the question. The more students are in the class, the more diverse answers could appear and the longer scoring time will be needed.

An automated grading system can assist the lecturer by reducing the grading time and enhance the learning process. Spending less time in grading enable the teacher to deliver faster feedback so that both of the teacher and the students can discover which aspect that the students have to improve. The lecturer can also think of another way to help the students in gaining a better understanding of the course.

Some researches and commercial options are proposed to discover a better solution to grade open questions automatically. However, the existing researches and commercial options are only suitable for or are only applied to a limited type of questions while there are various types of open questions. Therefore, this study aims to investigate how a question or an answer could be formulated to simplify the work of an automated grading system.

1.2. Problem Definition

Several studies have been conducted in this field. Each researcher applies different methods to their own system. However, some of the researches are not available anymore or cannot be accessed. On the other side, most of the researches explore on how to assess short-answer questions, which requires an answer of one phrase to one paragraph and maximally 100 words (Burrows, Gurevych, & Stein, 2015; Pribadi, Permanasari, & Adji, 2018), or an essay, which is graded based on several attributes, such as development, word choice or grammar usage, and organization (K. Zupanc & Bosnić, 2015). Occasionally, an open question requires answer written in more than one paragraph, but less than an essay, or even in a table and picture form.

In addition, most of the researches for automated grading utilize the existence of enormous training samples. The course with a large number of students has the privilege of providing huge data. However, a course with a limited amount of students has a limited size of sample data.

Various commercial options of the automated grading system are available in the market, for example, ETS, Gradescope, and CODAS. Several Learning Management System (LMS), such as Moodle, Populi, and Bookwidgets also implement automated grading within their system, but most of them can mark short-answer questions with limited capability. For example, Populi can grade the answer which matches exactly with the key answer; the non-matching answer should be graded manually. Gradescope can reduce teacher's burden by grouping similar answers based on defined rubric so the teacher can evaluate the answers faster. However, Gradescope is more suitable for engineering or mathematical subjects rather than analytical or conceptual subjects. CODAS could grade an exam after the lecturer graded several exams beforehand so that those exams are used as the model answer, but the best result is achieved when the number of model answers is at least 50.

The LMS options are not suitable for the university since students could write different answers and different length of answers. Additionally, each university already have their own LMS – University of Twente, for example, have Canvas – and it is not possible to add another one only for the automated grading feature. Furthermore, deploying commercial options might be expensive and its benefit could not be perceived immediately.

Therefore, it is important to find out methods to grade students answer automatically for a different type of open questions, in various style and response with an inadequate number of data beforehand, less than 50 data or even none. Once the methods are discovered, they are advantageous to help teachers to formulate the question better so the student can answer in a more beneficial manner for the automated grading system. For that reason, this study has explored several approaches to grade two common types in open question exam, so that the question and the answer can be constructed well.

1.3. Research Question

The research question for the master thesis is.

RQ: How can text mining and machine learning techniques be used for the automated grading of open questions?

The research question can be defined in the following sub-questions.

SQ1: How should open questions be formulated to be useful for automated grading using text mining?
An open question can have many possibilities of different answers. A clear instruction on the question can help the student to write reasonable answers and might simplify the grading process, so the reliability of the automated grading is enhanced. This sub-question can be answered by choosing several types of open question that are commonly used in an exam. The answer to these questions is processed by the system. Then, the system result is compared with the real result and analyzed to determine useful characteristics of a question for an automated grading system.

SQ2: What kind of text mining and machine learning techniques are available to grade open questions?
The current trend shows that information extraction, which is a part of the text mining technique, and machine learning are the most common techniques in automated grading. This study focuses on these techniques. To solve this question, a literature review that emphasizes these techniques was conducted.

SQ3: How to design an algorithm based on text mining and machine learning techniques to support the automated grading of open questions?
After acquiring the knowledge of some techniques in text mining and machine learning, a design of an algorithm based on the knowledge can be created. Then, a prototype was made to grade answers for open questions.

SQ4: In what way can the system performance be measured?
The performance of the prototype is essential to know how good the design is. An experiment was created to examine the performance, and suitable measurement is selected. The result of the validation was analyzed to determine the performance. Furthermore, an evaluation meeting is conducted to receive feedback from lecturers about the algorithm.

1.4. Research Methodology

This study uses the Design Science Research Methodology (DSRM) framework from Peffers et al. (2008). Figure 1 below illustrates the process model of the framework (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2008). This framework is chosen because the process is suitable to guide the process of the current study. The study started with problem identification and solution definition after conducting a systematic literature review in automated grading field. This step is explained in Chapter 1 and Chapter 2. Then, an artifact is created to solve the problem. The design of the artifact is described in Chapter 3. Next, the artifact is tested in an experiment and evaluated by potential users to measure the performance and to receive feedback. The experiment design, the results, and the conclusions are elaborated in Chapter 4 until Chapter 6. Finally, the result of the study is presented in a public defense.

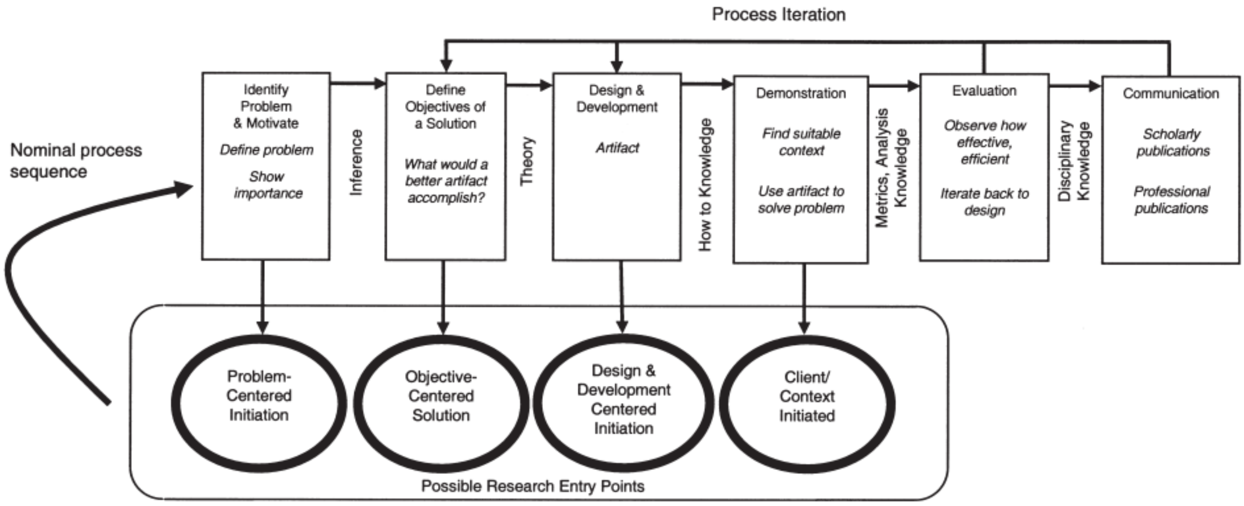

Process Iteration
Nominal process sequence
Identify Problem & Motivate
Define problem
Show importance
Inference
Define Objectives of a Solution
What would a better artifact accomplish?
Theory
Design & Development
Artifact
How to Knowledge
Demonstration
Find suitable context
Use artifact to solve problem
Metrics, Analysis Knowledge
Evaluation
Observe how effective, efficient
Iterate back to design
Disciplinary Knowledge
Communication
Scholarly publications
Professional publications
Problem-Centered Initiation
Objective-Centered Solution
Design & Development Centered Initiation
Client/ Context Initiated
Possible Research Entry Points

Figure 1 DSRM Process Model by Peffers

The explanation of each process is described below.

1. Identify Problem and Motivate
   This study was begun by identifying the problem. The problem was discovered after doing a systematic literature review (SLR) to explore current trends of the automated grading system.
2. Define the Objectives of a Solution
   After the problem was found, the objective of the study was defined. The problem can be solved by finding proper approaches for grading different types of open question automatically. As a result, different questions and techniques should be investigated to determine whether it is suitable or not for an automated grading system.
3. Design & Development
   In this phase, an artifact was made, based on the SLR result, as an embodiment from the solution. The measurement metrics for the evaluation phase was also decided.
4. Demonstration
   After the design and development phase, the artifact was tested on actual student answers in the final exam from two courses in Business and IT program: e-commerce and Business Case Development for IT. The student answers were anonymized before the experiment. Then, the result of the testing was evaluated and analyzed in this step to examine the performance of the artifact. If the result was not satisfactory, the design of the artifact should be altered until the acceptable result is achieved.
5. Evaluation
   After the experiment was done, there was an evaluation meeting with several lecturers in the University of Twente to present the artifact and ask their opinion about the artifact. The evaluation adopts the Unified Theory of Acceptance and Use of Technology (UTAUT) by Venkatesh et al. (2013) (Venkatesh, Morris, Davis, & Davis, 2003).
6. Communication
   The final part is to present the result of the study in a public defense.

1.5. Contributions

Previous works in automated grading system focus on various techniques to create a better system to grade an answer to open questions automatically. Essays and short answers are typical datasets used in this field of study. On the other hand, open questions in higher education are not limited to those types, such as give the characteristics of a concept and explain it in examples in real life, draw a picture, or fill in the table.

This part discusses the contributions of the current study, both to the theory and to the practice in the education system.

1.5.1. Contributions to Theory

This research provides several contributions to the theory as follows.

1. The thesis examines useful features of two types of open questions and answers so that it can be graded easier by an automated system. Each type of question has their own properties, and one method cannot be applied to all of them. Consequently, other methods are required to evaluate the answer of other open question types. To the author's knowledge, no studies were investigating this field.
2. The thesis proposes methods for automated grading system using combinations of text mining and machine learning for the small dataset. Former researches tend to select one technique between text mining and machine learning to build their system. Moreover, they have an enormous number of pre-scored answers as the dataset. Current work combines text mining and machine learning to examine the performance of these approaches for the small and ungraded dataset.
3. The thesis presents the use of RapidMiner for automated grading implementation. Previous works in the automated grading system rarely mention the tools they used. The current study probably is the first one to implement an automated grading using RapidMiner.

1.5.2. Contributions to Practice

Additionally, the current study offers several benefits for lecturers and students.

1. The thesis discovered the valuable properties of several types of open question. These properties are beneficial to build a suitable method for an automated grading and could lead to other benefits.
2. The thesis identified good practices to assist lecturers in creating a clear and comprehensible question so that the students can write their answer in the desired manner. Using these practices could reduce the workload of the lecturers in grading their exams because the students' answers are given in a similar format, which in turn means little variation in terms of style.
3. The thesis also provides some recommendations for questions and answers format. There are diverse characteristics of each open question. A question that asks students to mention and explain a few of examples differs with a question that asks students' perspective on a topic. The recommendations contain various suggestions for the lecturer to create a test that can be scored automatically.

CHAPTER 2 – LITERATURE REVIEW

This chapter consists of several theories used in this study from various articles. Section 2.1. explains about open questions and the level of intellectual understandings in a question. Next, section 2.2. and 2.3. describes several techniques in machine learning and text mining. Then, section 2.4. discusses the trends in automated grading based on the work of Burrows et al. (2015). Additionally, section 2.5. presents the datasets used in previous researches in automated grading. After that, section 2.6. elaborates the relationship between machine learning, text mining, and automated grading. Finally, section 2.7. discusses several common tools used in data mining.

This research used the Systematic Literature Review (SLR) as the method to find relevant literature. Work from (Rouhani, Mahrin, Nikpay, Ahmad, & Nikfard, 2015) is used as a guideline to do the SLR. The SLR was begun with the search process in scientific databases. The following digital libraries were selected to perform the SLR process because they provide broad coverage and highest impact full-text journals and conference proceedings (Rouhani et al., 2015).

- Scopus (https://www.scopus.com/)
- ACM Digital Library (https://dl.acm.org/)
- ScienceDirect (https://www.sciencedirect.com/)
- Google Scholar (https://scholar.google.com/)

There were several keywords used to find the relevant studies through the title, abstract, and keywords: ("automat* grading" OR "automat* scoring" OR "automat* assessment" OR "text grading" OR "text scoring" OR "machine learning" OR "text mining" OR "Natural language processing") AND ("open-ended" OR “open question” OR essay OR "short answer") AND (test OR evaluat* OR exam OR question).

From the search result, this study used several inclusion and exclusion criteria to help in selecting relevant papers. The inclusion criteria in this study are:

- Published between 2008 – 2018
- Studies related to the automated grading system, machine learning, text mining, short-answer assessment, open-ended assessment, or essay assignment evaluation

The exclusion criteria are:

- Studies that are not in English
- Studies that are done for an English assignment
- Duplicated sources of the same study (based on its title and abstract)
- Studies that are not related to the automated grading system, open-ended assessment, short-answer assessment, or essay assignment evaluation
- Studies that cannot be accessed, either because it must be purchased or because it is not available

Additionally, other than selected papers, several articles were included by looking at the references of a paper (backward search) or reviewing other articles which have cited a particular study (forward search) to obtain more information and clearer understanding (Levy & Ellis, 2006).

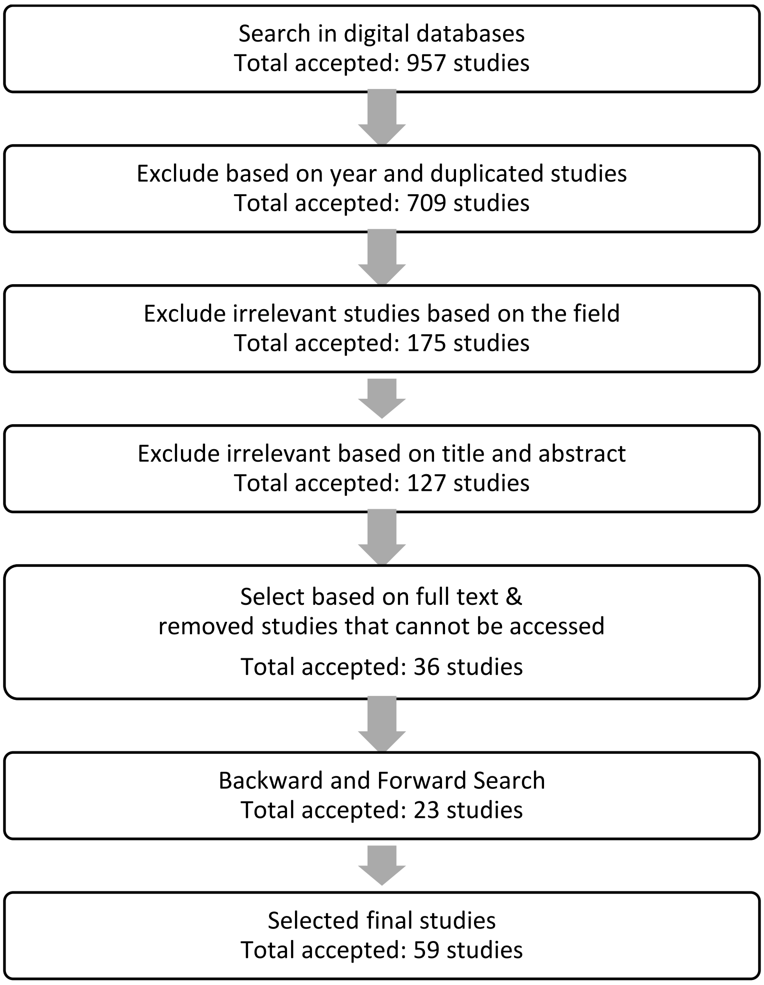

Search in digital databases
Total accepted: 957 studies
Exclude based on year and duplicated studies
Total accepted: 709 studies
Exclude irrelevant studies based on the field
Total accepted: 175 studies
Exclude irrelevant based on title and abstract
Total accepted: 127 studies
Select based on full text &
removed studies that cannot be accessed
Total accepted: 36 studies
Backward and Forward Search
Total accepted: 23 studies
Selected final studies
Total accepted: 59 studies

Figure 2 SLR Process

2.1. Open Questions

Open questions are a common way to evaluate student's understanding and knowledge in a topic is (Gonzalez-Barbone & Llamas-Nistal, 2008). Students are free to construct their ideas, concepts, and thoughts into the answer; hence the variety of student answers is inevitable. The answers given by the students depend on the way they perceive the question, so the composition of the question has an important role. An unambiguous and comprehensible question is preferred to get the students to understand what answer they should write (Husain, Baisb, Hussain, & Samad, 2012).

Bloom's Taxonomy of Educational Objectives is a common standard used for teachers to design their learning process, including when creating an exam. The taxonomy was found in 1956 by Dr. Benjamin S. Bloom after the 1948 convention of the American Psychological Association decided a classification of the understanding level that students can obtain in a class would be helpful (Clay, 2001). There are three education domains mentioned in Bloom's Taxonomy: cognitive, affective, and psychomotor domain (Ahmad, Adnan, Abdul Aziz, & Yusaimir Yusof, 2011). The cognitive domain related to intellectual skills, which involve the ability to recall what had been learned previously and very useful to test student's understanding of the particular topic. The affective domain includes how people cope with things emotionally, such as feelings, values, appreciation, enthusiasms, motivations, and attitudes. The psychomotor domain consists of physical movement, coordination, and use of the motor-skill area.

Written exam tends to ask questions in the cognitive domain. Meanwhile, the affective domain is mostly used in group discussions or project assignments where students can learn all aspects in this domain through group dynamics in teamwork. Then, the psychomotor domain applies to practical tests. Because this research is about automated grading system for written assignment, only cognitive domain is explained more.

There are six levels of intellectual understanding in a cognitive domain based on their order (Ahmad et al., 2011; Clay, 2001; Patil & Shreyas, 2017; Sangodiah, Ahmad, & Ahmad, 2014):

a. Knowledge: to remember, recognize, and recall previously learned material, such as dates, events, definitions, theories, and procedures. The keywords used often are choose, define, describe, find, identify, inquire, know, label, list, match, memorize, name, outline, recall, recognize, reproduce, select, and state.
b. Comprehension: understanding the meaning of information, then translating, interpreting, and explaining it again in students' own words. It can also cover predicting outcome and effects, classify, or compare. Common keywords used in this level are compare, comprehend, convert, defend, demonstrate, distinguish, estimate, explain, extend, generalize, give examples, infer, interpret, paraphrase, predict, restate, rewrite, summarize, and translate.
c. Application: invokes student's ability to apply learned material, such as general rules, methods, or principles, in a new situation to solve a problem. Several keywords for this level are apply, build, change, compute, construct, demonstrate, discover, illustrate, manipulate, modify, operate, plan, predict, prepare, produce, relate, show, sketch, solve, and use.
d. Analysis: breaking down concepts and components into smaller units, identify the relationship between the components and with the overall concepts. A few keywords that belong to this domain are analyze, assume, break down, categorize, classify, compare, contrast, diagram, deconstruct, differentiate, discriminate, distinguish, experiments, identify, illustrate, infer, outline, relate, select, separate, subdivide, and test.
e. Synthesis: putting elements altogether to create a new functional whole product. A few keywords of this domain are categorize, combine, compile, compose, create, design, develop, devise, explain, generate, modify, organize, plan, rate, rearrange, reconstruct, relate, reorganize, revise, rewrite, score, select, summarize, tell, and write.
f. Evaluation: using evidence, standards, and reasoned argument to create judgments of differences, controversies, or performance of a design. Some keywords in this domain are appraise, compare, conclude, contrast, criticize, critique, decide, defend, describe, discriminate, evaluate, explain, interpret, justify, relate, summarize, and support.

In the higher-education level, the lecturers usually use open-question-based examination because it can assess the depth of student's understanding in the class. A good question is unambiguous and comprehensible to the students. Moreover, a good exam consists of a set of questions that ask different levels of understanding to facilitate the students developing critical thinking ability (Ahmad et al., 2011). Therefore, it is important for the lecturers to formulate questions that examine more than one level of intellectual understanding in an unambiguous and comprehensible way.

2.2. Machine Learning

Machine learning has become a popular technology in recent years. Machine learning is a study of how to use computers to simulate human learning activities (Lv & Tang, 2011). Unlike human learning that uses memory, thinking, perception, feeling, and other mental activities, machine learns using the knowledge and skills gained from the environment (H. Wang, Ma, & Zhou, 2009). Another focus of machine learning is how to improve the performance of the learning process. Several applications, such as spam filtering, recommender system, and face recognition, are implemented using machine learning methods. Machine learning enables the system to perform a task by finding patterns and learning from experience, which is provided by large amounts of data (Ivanović & Radovanović, 2015; Kwok, Zhou, & Xu, 2015).

There are four common types of machine learning techniques: supervised, unsupervised, semi-supervised, and reinforcement learning (Ivanović & Radovanović, 2015; Kwok et al., 2015; Lee, Shin, & Realff, 2018). Supervised learning utilized labeled examples to derive patterns and apply the patterns into new examples. On the other hand, unsupervised learning deals with unlabeled data to learn about the relationships between examples and group them based on their relationships. Semi-supervised learning combined labeled and unlabeled data to assist the supervised learning. Reinforcement learning is related to the learning algorithm within an entity, such as a software agent or robot, to decide what actions they have to do based on the input data from the environment.

Supervised learning works with labeled data, often called training data. The labeled data contain the value for each information and its corresponding class/category. This data is used in the training process to build a model so that it can determine the pattern or conclusion from the data and apply it to the new data. To conclude from training data, having useful and meaningful features extracted is important (Kwok et al., 2015).

Besides training data, there is also testing or evaluation data. After performing the learning process and building a classifier model, it is important to evaluate the performance of the model. The performance evaluation is done by applying the model to the new data. To ensure the model performance and avoid any bias, the evaluation data should be different with the training data (H. C. Wang et al., 2008).

Gaining more training data tends to produce a more confident model. However, not all institutions could acquire a huge amount of sample data. To overcome the limited number of data, a common practice is to perform cross-validation technique (Hladka & Holub, 2015; H. C. Wang et al., 2008). The main idea of cross-validation is to divide the data into *k* subsets of equal size. At the *i*-th step of the iteration, the *i*-th subset is used as a testing data, while the remaining parts form the training set. Therefore, all data partitions act as training and testing (for once) dataset in *k* number of iterations.

Unsupervised learning does not have labeled examples. Therefore, there is no training and testing data in unsupervised learning. The aim of unsupervised learning is finding relationships between the data and grouping them without any outside information, but on the intra-similarity and inter-similarity between examples (Ivanović & Radovanović, 2015; Khan, Daud, Nasir, & Amjad, 2016). Clustering is the most common technique in unsupervised learning.

2.3. Text Mining

Text mining is a process that applies a set of algorithms to extract interesting patterns from textual data sources and analyses the patterns to gain knowledge and facilitate decision making (Aggarwal & Zhai, 2013; Dang & Ahmad, 2015; Talib, Hanify, Ayeshaz, & Fatimax, 2016). Text mining is related to other fields, such as data mining, web mining, machine learning, statistics, information retrieval, information extraction, computational linguistics, and natural language processing (NLP) (Dang & Ahmad, 2015; Talib et al., 2016).

The techniques of text mining include information retrieval, information extraction, text summarization, text classification, and clustering (Aggarwal & Zhai, 2013; Agrawal & Batra, 2013; Dang & Ahmad, 2015; Talib et al., 2016; Vijayarani, Ilamathi, & Nithya, 2015):

a. Information Retrieval

Information Retrieval (IR) is a task of extracting relevant and associated information from a collection of several resources according to a set of given words or phrases. The most well-known IR systems are search engines, such as Google and Yahoo, which identify related documents or information based on the search query. The search queries are used to track the trends and attain more significant results so that the search engine produces more relevant and suitable information to user needs.

b. Information Extraction

Information extraction (IE) is a method in text mining to identify and extract meaningful information, such as the name of a person, location, and time and relationships between the information within the text. The extracted corpus is in structured form and stored in a database for further processing to get the knowledge inside the text. This method is advantageous when handling huge volumes of text.

c. Text Summarization

Text summarization is a task of generating a concise version of the original text. The source of original text can come from one or multiple documents on a particular topic. The summary contains important points of the original document(s). During the process of producing a coherent summary, several features, such as sentence length, writing style, thematic word, and syntax, are considered and analyzed.

d. Text Classification

Text classification, or text categorization, assigns a natural language document into categories based on its content. The current approach in this procedure is to train pre-classified documents using machine learning. The pre-defined categories are symbolic labels with no additional semantics. The goal of text categorization is to classify a new document into one or more group, depends on the context, or to rank the categories by their estimated relevance to the document.

e. Clustering

Clustering groups similar documents without any pre-defined label, hence training data is not required. In a cluster, similar terms or patterns extracted from the documents are grouped together. Good clustering technique groups similar objects in the same cluster, while objects from two different clusters are dissimilar.

Text is an unstructured data. Preparing the data beforehand is important to obtain knowledge from text. Several typical pre-processing steps for textual data are (Omran & Ab Aziz, 2013; Quah, Lim, Budi, & Lua, 2009; Shehab, Elhoseny, & Hassanien, 2017; Vijayarani et al., 2015):

a. Tokenization

Tokenization divides the text into sentences or individual words (tokens). The delimiter for this process could be non-letters characters, such as whitespace or punctuation symbol.

b. Stop words removal

Stop words are common words that are unnecessary and do not affect the main idea of a text if it is removed. Removing stop words reduces the dimensionality of the term space and retain important words so those keywords can be used for further analysis. Commonly used stop words in the English language include the auxiliary verbs and the preposition question words, such as a, the, is, with, to, at, an, what, where, that, etc.

c. Stemming

Stemming is used to identify and trim words to its root/stem form, by removing prefixes and suffixes. For example, the words consider, considers, considered, and considering can be trimmed to the word “consider." The purpose of this technique is to reduce the total number of words without removing the essence of the text.

d. Generate *n*-gram

An *n*-gram is a sequence of adjacent *n* character or word in the text. An *n*-gram of size one is called as a unigram, size two is a bigram, and size three is trigram. For example, in sentence “I came late today”, the unigram is “I”, “came”, “late”, and “today”; the bigram produces “I came”, “came late”, “late today”; and the trigram are “I came late” and “came late today”.

After pre-processing is done, the features of the text can be created. For text data, the features are represented in one of the word vectors, such as Term Frequency or Term-Frequency-Inverse Document Frequency (Bafna, Pramod, & Vaidya, 2016; Manning, Raghavan, & Schütze, 2009; Vijayarani et al., 2015). The word vectors transform text into more structured data and can be understood easier by the computer in the form of Term Document Matrix (TDM).

a. Term Frequency (Driscoll et al.) (Driscoll et al.) is a value between a word *w* and a document *d*, based on the weight of *w* in *d*. The weight of TF is equal to the number of occurrences of word *w* in document *d*.
b. Term-Frequency-Inverse Document Frequency (TF-IDF) is a numerical statistic that represents a word importance to a collection of documents. The TF-IDF value increases proportionally to the number of times the word occurs in a document, but is counterbalanced by the frequency of the word in the collection. The value of TF-IDF is the highest when word *w* appears many times within a small number of documents; the lowest when the word occurs in all documents; lower when the word appears fewer times in a document, or in many documents.

2.4. Automated Grading

The researches in the automated grading system were started by Page in 1966 with the Project Essay Grading (Wresch, 1993). Twenty-five years later, there was little interest in using a computer to grade an essay (Wresch, 1993). However, the number of researches in the automated grading system in the last decade is increasing as this field contains opportunities and possibilities to be explored more. The classification of each study found is based on Burrows et al. (2015) themes. Although Burrows classification is about automated short answer grading, it is also applicable to other open question assignments.

Burrows et al. (2015) did a historical analysis through researches in automated short answer grading (ASAG) system and determined five temporal themes along the researches, which are the era of concept mapping, the era of information extraction, the era of corpus-based methods, the era of machine learning, and the era of evaluation. In concept mapping, the basic idea is considering student answers as several concepts to be checked later about its existence in the grading process. Meanwhile, information extraction is a series of pattern matching techniques that can extract structured data from unstructured sources to find any fact related to the answer. Corpus-based methods utilize statistical properties of large document corpora which can be used to detect the synonyms in an answer and prevent misinterpretation of similar correct answers. On the other hand, machine learning techniques employ measurements extracted from NLP approaches and similar to be combined later into one score using a classification or regression model. At last, the evaluation era is not method related: they use shared corpora, also competitions and evaluation forums between research groups around the world on a particular problem for money or prestige.

Table 1 Previous Works in Automated Grading System

Theme	Works by
Concept Mapping	Wang et al., 2008; Jayashankar & Sridaran, 2017
Information Extraction	Siddiqi & Harrison, 2008; Sima, Schmuck, Szöllosi, & Miklós, 2009; Lajis & Azizi, 2010; Cutrone, Chang, & Kinshuk, 2011; Gutierrez, Dou, Martini, Fickas, & Zong, 2013; Omran & Ab Aziz, 2013; Srivastava & Bhattacharyya, 2013; Jayashankar & Sridaran, 2017; Mehmood, On, Lee, & Choi, 2017; Pribadi et al., 2018
Machine Learning	Wang et al., 2008; Bin, Jun, Jian-Min, & Qiao-Ming, 2008; Ziai, Ott, & Meurers, 2012; Gutierrez et al., 2013; K. Zupanc & Bosnic, 2014; Nedungadi, L, & Raman, 2014; Rahimi et al., 2014; Wolska et al., 2014; Dronen, Foltz, & Habermehl, 2015; Jin & He, 2015; Kudi, Manekar, Daware, & Dhatrak, 2015; Phandi, Chai, & Ng, 2015; Nakamura, Murphy, Christel, Stevens, & Zollman, 2016; Wonowidjojo, Hartono, Frendy, Suhartono, & Asmani, 2016; Latifi, Gierl, Boulais, & De Champlain, 2016; Perera, Perera, & Weerasinghe, 2016; Jin, He, & Xu, 2017; Mehmood et al., 2017; Shehab et al., 2017; Zhao, Zhang, Xiong, Botelho, & Heffernan, 2017
Corpus-based	Vajjala, 2018

Based on Table 1, the most popular theme is machine learning and followed by information extraction. There is no result from evaluation era because there is no report mentioned about this. However, several studies use the same dataset retrieved from the same source which is from Automated Student Assessment Prize (ASAP) competition by Kaggle, especially for automated essay grading, but they do not compete each other. They use the similar dataset as it is publicly available or they want to compare the performance of their system with the others that use the same dataset.

One system is not always associated with one theme only, for example, the superlative model (Jayashankar & Sridaran, 2017) and hybrid ontology-based information extraction (Gutierrez, Dou, Martini, Fickas, & Zong, 2013) system. Several systems based on machine learning are also included as information extraction because the pre-processing phase in machine learning method extract some features of the text before the technique can process the answer. Since machine learning and information extraction are prevalent methods among the other, this study focuses on several types of research using these two methods.

Information Extraction

Information extraction (IE) aims to gather relevant facts or ideas in a text answers, either explicitly stated or implied, by applying a set of patterns (Hasanah, Permanasari, Kusumawardani, & Pribadi, 2016; Roy et al., 2015). The patterns are employed in the words, phrases or sentence level, syntactically or semantically. The evaluation in IE techniques is usually done by matching the patterns, that are found in the training dataset or defined by the human grader, with the answers to be graded. Several typical techniques in this era are parse trees, regular expression matching, syntactic pattern matching, and semantic analysis.

Jayashankar and Sridaran (2016) presented an IE-based method on word-level matching. Their model breaks the answers into keywords, which are represented by two different word clouds named cohesion and relative cloud. Cohesion cloud contains common words between student answer and the answer key, while uncommon words are included in the relative cloud. Then, the teacher will evaluate the answer by counting the number of words in the cohesion cloud and mark the answer. The agreement rate for this model was 98%, and the accuracy score deviation from the mean was 2.82. The agreement rate shows the promising result as it achieves nearly perfect agreement with human scoring. The accuracy score deviation is one factor to assess the efficiency of any automated short answer analysis tool besides cost and time taken. The lower the value is, the more efficient the system is. The deviation score of the superlative model was lower than IndusMarker, the latest development of automated grading system at the moment, which means the superlative model has better performance than IndusMarker.

The works of Omran and Aziz (2013) and Pribadi et al. (2018) performed sentence-level similarity in their system. The system requires a model answer to be compared with the student answers based on the similarity. They also utilized the Longest Common Subsequence (LCS) within the system to calculate the most accurate sequence by counting the letters in the sentence as one whole string (Omran & Ab Aziz, 2013). The differences are the matching process and the scoring method. Omran and Aziz generated a large number of sentences for the model answer to cover all possible answers by rewriting the model answer in its synonym. In each phase of the answer processing, which used common words and semantic distance beside LCS, Omran and Aziz assigned a score and combined all of them by weighting the smooth factor. On the other hand, Pribadi et al. compared the students answer with lecturer answer, to find which student answers were the closest matches to the lecturer answer, using Maximum Marginal Relevance (Joiner et al.) (Joiner et al.) method. Pribadi et al. graded the answer based on its similarity with the reference answer using the geometric average normalized-longest common subsequence (GAN-LCS) technique. The evaluation result of both works shows a satisfactory result. The method by Omran and Aziz obtained Pearson *r* value is around 0.80 – 0.82 and the system performs better than the Latent Semantic Analysis (LSA) technique. Meanwhile, Pribadi et al. achieved an average accuracy of 91.95% in generating the reference answer variation, the correlation value of 0.468, and root mean square error (RMSE) value of 0.884. MMR method accepted a reference answer candidate if the score was equal to or more than four and rejected one if the score was lower. The system accepted 240 out of 261 correct answers. Thus it scored 91.95% correctly. Compared to other works that use the same dataset, the RMSE value of this study is the best one, and the correlation value is the third best.

Other techniques in IE are syntactic pattern matching and semantic analysis. Syntactic pattern matching uses syntactical structures from the model answer to grade the student answer, by chunking the text, parsing, part-of-speech (POS) tagging, sentence segmentation, syntactic templates, tokenization, or word segmentation (Burrows et al., 2015). Srivastava and Bhattacharyya (2013) and Siddiqi and Harrison (2008) developed a model to evaluate short answer questions based on syntactic pattern matching. Semantic analysis, which is used in Auto-Assessor by Cutrone et al. (2011), focuses on finding the similar meaning, usually from the synonym, of the answer with the model answer.

Captivate Short Answer (CSA) evaluator by Srivastava and Bhattacharyya (2013) operates in two modes which are automatic, the mode that requires minimal human effort because the system generates the scoring model automatically, and the advanced mode where examiner can tune and customize components of the scoring model. The evaluation is done by evaluating 30 responses about Class-7 General Science questions using automatic and advanced model. The correlation coefficient of the advanced model is higher than auto model because advanced model enables assessor to review the automatic extracted features, select relevant synonyms and phrases, specify multiword concepts, and define advanced scoring logic, which improves evaluation accuracy.

Siddiqi and Harrison (2008) developed a prototype system to mark short-answer answers automatically. The system process answers from undergraduate biology exam at The University of Manchester in three steps: spell checking and correction, parsing, and comparison. The comparison process compares the tagged and chunked text from previous steps with the required syntactical structures, that are constructed in Question Answer Language (QAL). The system also compares any grammatical relations in the student answer with the examiner-specified grammatical relations. After the comparison is made, the result goes to the marker to give the final score. The performance is measured in human-system agreement, and the result was 96%, which is excellent and higher than other IE-based systems on previous works. However, the dataset in those previous works is different while it is mandatory to use the same dataset to obtain an effective comparison.

Another prototype system was also created by Cutrone et al. as a Windows application. The system emphasizes on WordNet processing to match the words exactly based on matching on POS tag, the word match, and the words, that have been matched, have an equivalent relative position in the sentence concerning the sentence verb(s) (if any exist). There are three different user roles implemented in the system: the Assessor, the Student, and the Operations personnel. The Assessor role creates the test, the Student takes the test and can review the scores, and the Operations initiate the system to grade the test. Because the system focuses on the single-sentence response, which is free of grammar and spelling mistakes, the assessor and student are expected to input the answer in without grammatical or spelling error. The system performance is observed from the agreement level and total grading time spent. Unfortunately, there is no data about the evaluation result.

Sima et al. (2009) introduced "answer space," the formal description to define a set of answer types, syntactic structures, and possible grammatical structure constructors, as a method deployed in eMax system. In eMax, student answers are examined in three main steps: syntactic analysis, semantic analysis, and scoring. Syntactic analysis phase will check student and teacher answers. If there is no match for the student answer, the system will mark the answer to be manually assessed. During the manual assessment, an additional feature of this eMax version in which the answer space can be updated if it is needed. If a match is found, the answer will go to the scoring phase. In some cases, when there are two matches found between student and teacher answer, semantic analysis will determine the closest match. The answer will be graded based on the closest match. After applying the system to the real examinations, 72% of the answers were graded automatically, where 7% of them was scored incorrectly, and 28% needed a manual review from the lecturer. After the review, 17% of the manual assessment obtained the same mark as the eMax, 11% had to be corrected by the professor. The results show pretty good accuracy and the additional feature improve the eMax performance.

Auto-Assessor and CSA evaluator are more likely about automated grading in an e-learning system. Both of them does not mention about using particular dataset because the question and answer key are submitted by the teacher through the system and the similarity between students' answers and answer key is compared. Unfortunately, there is no detail data about the experiment result of Auto-Assessor, but only an explanation and it is hard to follow when there is no data.

Superlative model and eMax does not leave out the role of the lecturer in grading completely. The systems are tools to help the lecturer to grade the answers more efficiently. In the superlative model, the lecturer grades an answer based on the word cloud generated by the system. eMax reduces the grading load by delivering the rejected answers to the professors so they can review it and updated the model answers.

The performance of IE-based systems is mostly satisfactory because most of the correlation or agreement rate value are above 80%. The table below display the summary of previous works in an automated grading system based on IE methods.

Table 2 The Summary of Previous Works based on Information Extraction Methods

Work of	Year	System / Method name	Theme	Dataset	Assignment for Evaluation	Evaluation Method	Measurement and Result
Siddiqi & Harrison	2008	N/A	Information extraction: syntactic pattern matching	Undergraduate biology exam at The University of Manchester	Testing set of the dataset	Grade the unseen testing set	Human-system agreement: 96%
Sima et al.	2009	eMax	Information extraction: syntactic & semantic analysis	N/A	Computer Architectures tests	Random sampling and comparison of evaluation results	Accuracy
Cutrone et al.	2011	Auto-Assessor	Information extraction: semantic word matching	N/A	Questions & student answers	Comparing the grade of the system and human markers	Agreement and scoring time
Srivastava & Bhattacharyya	2013	CSA evaluator	Information extraction: syntactic analysis	N/A	12 different answers from 30 questions in Class-7 General Science	Validate the system to score the assignment	Correlation Coefficient Auto Model: 0.66 Advanced Model: 0.81
Omran & Aziz	2013	Alternative Sentence Generator Method and text similarity matching	Information extraction: sentence-level similarity	Pre-scored assignments from introductory computer science assignments of undergraduate students	Testing set of the dataset	Compare the system with human marking, other automated grading systems, and other technique	Correlation Measurement with Human Grade: 0.80-0.82 Correlation Measurement with another system: 82%
Jayashankar & Sridaran	2017	Superlative model using the word cloud	Concept mapping, information extraction: word-level matching	Student responses and answer key	IGCSE board examination for Grade X	Compare the system with human marking	Agreement rate: 98% Accuracy score deviation from mean: 2.82

Pribadi et al.	2018	MMR and GAN-LCS	Information extraction: sentence-level similarity	Pre-scored Texas Corpus	Pre-scored Texas Corpus	Grading the assignment with the proposed method and evaluate the result	Accuracy: 91.95% The correlation value: 0.468 RMSE value: 0.884

Machine Learning

Various automated grading systems implement different machine learning algorithm, and the most prevalent techniques are classification and regression (Burrows et al., 2015; Roy et al., 2015). In this study, the most popular algorithms deployed are Support Vector Machine (SVM). Other algorithms are Latent Semantic Analysis (LSA), Naïve Bayes, k-Nearest Neighbors (KNN), Neural Network, and Random Forest.

One algorithm can be combined with another algorithm to enhance the performance, instead of using only one algorithm, but one algorithm can be compared with the others to discover which performs better than the other.

Support Vector Machine (SVM)

Research by Wang et al. (2008) created and compared three automated grading methods based on the availability of concept identification in the system and how the system grades the answer. The three methods were pure heuristics-based grading (PHBG), data-driven classification with minimum heuristics grading (DCMHG), and regression-based grading (RBG). PHBG technique identifies the concept by representing text objects as word vectors. PHBG uses TF-IDF weighting scheme as the metric, and the grading is done by mapping the answers to numeric scores using the prescribed scoring heuristics. DCMHG performs concept identification by categorizing the text using the SVM method, and the grading is executed similarly as PHBG. RBG does not operate any concept identification, and the grading is conducted using SVM regression. Cohen's kappa indicates the performance of concept identification and DCMHG achieves a better result than PHBG. Since RBG does not perform concept identification, the result for concept identification is produced for PHBG and DCMHG method only. The result of *r* value shows the highest reliability with human scoring is achieved by the DCMHG method over all methods. Overall, all three methods had satisfactory reliability (more than 0.80).

Lajis and Aziz (2010) utilized SVM to propose an approach called Node Link (NL), in which the expert and learner conceptual model are generated, the extracted terms from the answer are represented as a node, and each node is connected. Each node and the link between them are weighted to determine the score of the answer. The average of exact agreement of the system was quite low: 0.28, the exact or adjacent agreement was around 0.57, and the correlation value was 0.74. These values mean that the system does not score as similar as a human grader, but the consistency is pretty good. The system is also compared with other technique, such as Vector Space Model (VSM) and LSA, and the result shows the proposed technique performs better correlation result over the others.

Nakamura et al. (2016) implemented SVM and Naïve Bayes algorithm in an online tutoring system for introductory physics. The student can answer the question in around 1 or 2 complete sentences, and the system will give score and feedback about the answer. The Cohen's kappa result shows relatively good agreement between the performance of a self-validation on an entire response set and the performance of cross-validation on each half of the data set. Ziai et al. (2012) compared the performance of Comparing Meaning in Context in English (CoMiC-EN) System with previous work by Mohler et al. (2011). Because CoMiC-EN is a meaning comparison tool, the authors replaced the memory-based learning approach into a regression-capable learning strategy to do the scoring task instead of meaning comparison. The result of the correlation and RMSE value is lower than the previous work, but because the study aims to compare two systems from different research strands on the same dataset, the researchers assumed it does not matter.

Other applications of the SVM algorithm are used to find the semantic representations in an essay (Jin & He, 2015; Jin, He, & Xu, 2017). Jin and He (2015) examined the effect of three approaches, namely on-topic degree, Continuous Bag of Words (CBOW), and Recursive Autoencoder (RAE). On-topic degree and CBOW are features based on word vectors, while RAE is features based on sentence vectors. To evaluate the performance of these approaches, researchers used a baseline method which used four typical features used as indicators of essay quality. Then, the researchers explored the influence of each approach after it was added as an additional feature of the baseline method. The result of each approach shows the improved result for all measurements than the baseline. The experiment indicates that semantic features enhance the performance of automated grading machine. Another work of Jin, He, and Xu (2017) utilized semantic similarity features and distributed semantic representations of essays in automated essay scoring system. The study investigates the performance of semantic representations of essays for generating the semantic features for AES, which are Vector Similarity and Dimension Extension. The results show that the combination of best features generated by Vector Similarity and Dimension Extension could produce better performance for both algorithms used.

Latent Semantic Analysis (LSA)

Nedungadi et al. (2014) discussed Amrita Test Evaluation & Scoring Tool (A-TEST), a text evaluation and scoring tool that learns from course materials and human-rater scored text answers and also directly from teacher input. The system consists of 2 main phases: analysis and scoring phase. Analysis phase creates a Word by Context Matrix (WCM), which is taken from good and bad essays, and updated with LSA to reduce the dimension by removing unimportant details. Cosine similarity comparison is also made in this phase to prepare the scoring model. In the scoring phase, the essays are graded using the scoring model and multiple regression analysis. The experiment result is compared with two human raters. From the kappa value, the system shows a moderately good agreement with the human raters and adjacent agreement result obtain a high result, more than 95%.

Perera et al. (2016) created a system based on Vector Space Models (VSMs) and NLP techniques. The system took a set of student answers for short essays and relevant model answers to build the scoring model using LSA. The evaluation was done by comparing the result with the average score of human graders. The correlation result shows moderate to a strong correlated value between human and machine grader (around 0.67 and 0.813). However, the adjacent and exact agreement result is not good enough (most of them are below 50%). The researcher considered this is presumably caused by the inconsistency of the human grader, but there was no evidence for this.

The work of Wonowidjojo et al. (2016) observes the effect of syntactic information (in the form of word order) and Coreference Resolution in automated essay scoring using LSA. The researchers inspected two scenarios: document classification based on the cosine similarity measure between the answer essay and the essays in the training set, and not using classification, but calculating the average score based on the cosine similarities across all essays in the training set. The first scenario examined three variables: LSA vs. Syntactically Enhanced LSA (SELSA), Using vs. Not Using Coreference Resolution, and Maximum vs. Average Similarity. In the first scenario, LSA Average Similarity without Coreference Resolution obtained the highest result over all techniques and the same trend applied for the second scenario. On the other hand, SELSA worked better with Coreference Resolution in the first scenario and was quite stable in the second scenario. There is no explanation from the researchers the reason behind the result.

Naïve Bayes

Phandi et al. (2015) presented an approach of flexible domain adaptation for automated essay grading system using Bayesian linear ridge regression (BLRR). Domain adaptation is the task of adapting knowledge learned in a source domain to a target domain. The experiment used four set pairs of essays. The data for training and testing was divided randomly and into different sizes (10, 25, 50, and 100) in which the larger sets contains smaller sets. There were four configurations to examine the performance of BLRR by changing the ρ-value. The ρ-value is the correlation between latent scoring functions for the target and the source domain. The ρ-value is distributed from zero to one, where zero means a straightforward concatenation of the source and target data, and one is the shared hyperparameter setting, in which the Gamma distribution of the machine learning model is shared between source and target data. The different ρ-value used for each configuration are: SharedHyper have ρ=0, EasyAdapt have ρ=0.5, Concat have ρ=1, and ML-ρ have ρ maximizing the likelihood of the data. ML-ρ enable the model to traverse between three others configurations. Concat and ML-ρ mostly have the best result among other configurations. The result proves that domain adaptation is effective and important in the context of a small number of target essays with a large number of source essays as the Quadratic Weighted Kappa (QWK) is improved. A large number of source domain essay is not necessarily true because ten additional target domain essays can improve the result.

k-Nearest Neighbors (KNN)

Bin et al. (2008) utilized *k*-Nearest Neighbors (KNN) in their text categorization approach to scoring an essay automatically. The study inspects TF (Driscoll et al.) and information gain (IG) as the features selected for the algorithm. The features of the training sets were computed first. Then, the similarity of the test essays with all of the training essays was calculated using the cosine formula to search the k-nearest neighbors. The experiment shows the optimal result is achieved when *k*=3 or *k*=5. The maximum accuracy is obtained at 76%, which is satisfactory, using TF when *k*=5, the threshold is 40, and the argument is selected as the feature.

Others

Wolska et al. (2014) conducted a clustering-based approach to mark short responses to the computer-assisted test. Similar to other techniques, the participant responses were preprocessed. Then, *n*-gram, question material, and keyword-features were used as the selected feature of the preprocessed dataset. They used a single pass clustering method to group the answers. Because the purpose of the study is discovering the impact of scoring clustered responses to the efficiency, the evaluation was done by calculating the total scoring time per sheet divided by the number of responses on the given sheet (mean per response scoring time), the relationship between the amount of content per answer sheet, and the time it took to grade the content. The method shows positive results that grading clustered responses tend to be faster and is as efficient as marking the answer sheets manually.

Latifi et al. (2016) evaluated written responses for clinical decision-making (CDM) questions on a clinical competency examination using automated scoring. The framework is divided into three stages, which are extracting features from CDM responses, developing the automated scoring models, and scoring classification and error analysis. The second stage is done by training the computer model using a decision-tree induction algorithm by learning the characteristics of the extracted features from the first stage iteratively. The computer model is validated with the cross-validation method. The final step is marking the written responses, and the scoring accuracy is evaluated by comparing it with human marking. The agreement rate between human and computer model was 95.4%. The result shows almost perfect agreement between computer and human examiners based on Landis and Koch (1977). In addition, standardized mean score difference (SMSD) and *F* score (Gnimpieba, VanDiermen, Gustafson, Conn, & Lushbough) were also measured as distributional measurement, and the values were less than 0.15 (which meant the scoring model are interchangeable with human evaluators) and higher than 0.91 (which indicated the scoring models are indistinguishable from the human assessors) respectively.

The study process of machine-learning-based systems is similar to each other, from the pre-processing step, the dataset, the evaluation method, and the measurement metrics. The summary of the previous works based on machine learning methods is described in the table below.

Table 3 The Summary of Previous Works based on Machine Learning Methods

Work of	Year	System / Method name	Theme	Dataset	Assignment for Evaluation	Evaluation Method	Measurement and Result
Wang et al.	2008	PHBG, DCMHG, RBG	Concept mapping; machine learning: SVM; or both	Debris flow hazard (DFH) task	Testing set of the dataset	10-fold cross-validation	Cohen's Kappa (for PHBG and DCMHG method only) Pearson's *r* value: PHBG 0.90, DCMHG: 0.92, RBG: 0.86
Bin et al.	2008	Text Categorization Approach	Machine learning: KNN	Pre-scored CET4 essays of Chinese Learner English Corpus (CLEC)	Testing set of the dataset	Apply the method to training & testing dataset and evaluate the result	Precision and Recall
Lajis & Aziz	2010	Node Link (NL) scoring technique	Machine learning: SVM	Exams questions from various domains and categories of the university	Testing set of the dataset	Grade the testing set and compare it with other technique	Exact and adjacent agreement, Pearson correlation

				and school students			
Ziai et al.	2012	CoMiC-EN	Machine learning: SVMRanks and Support Vector Regression (SVR)	A corpus of ten assignments and two exams from an introductory computer science class	Testing set of the dataset	12-fold cross-validation	Pearson correlation: 0.405 RMSE: 1.016
Wolska et al.	2014	Clustering-based system	Machine learning: clustering	Placement tests for German as a Foreign Language	Placement tests for German as a Foreign Language	Grading the answers and measure the time	Mean per response scoring time and grading time, the relationship between the amount of content per answer sheet, and the time it takes to grade the content
Rahimi et al.	2014	Response to Text Assessment evaluation	Machine learning: Naïve Bayes, or Random Forest, or Logistic Regression	Pre-graded short essays written by students in grades 4–6	Testing set of the dataset	10-fold cross-validation	Accuracy, Kappa and QWK
Nedungadi et al.	2014	Amrita Test Evaluation & Scoring Tool (A-TEST)	Machine learning: LSA, multiple regression analysis	Pre-scored essays from kaggle.com, written by students from Grade 7 to Grade 10	Testing set of the dataset	Grade the assignment and evaluate the result	Kappa and QWK
Jin & He	2015	Latent Semantic Word Representations	Machine learning: SVMRank	Automated Student Assessment Prize (ASAP)	Testing set of the dataset	10-fold cross-validation	Normalized root mean squared error, Pearson's correlation coefficient, and Spearman correlation coefficient with the baseline
Phandi et al.	2015	Flexible Domain Adaptation	Machine learning: BLRR	ASAP dataset	Testing set of the dataset	5-fold cross-validation with a different sample of training data	QWK

Latifi et al.	2016	Technology-enhanced CDM framework	Machine learning: Decision-Tree Induction and J48 algorithm	Extracted features from CDM prescored responses from the online examination system of the MCC.	Ungraded CDM responses	Validate the system to score new response and evaluate the correlation	Agreement rate: 95.4% SMSD: 0.15 *F* score: 0.91
Nakamura et al.	2016	Automated analysis in an online tutoring system	Machine learning: classification using Naïve Bayes and SVM	Previous responses in the online system	No new data to validate the system	Self-validation using cross-validation	Cohen's kappa, Precision, and Recall
Wonowidjojo et al.	2016	Syntactically Enhanced LSA (SELSA) and Coreference Resolution	Machine learning: LSA	Pre-graded students essay assignment	Testing set of the dataset	Grade the assignment and evaluate the result	LSA without Coreference Resolution achieve the highest correlation than other techniques.
Perera et al.	2016	Vector Space Models and NLP techniques	Machine learning: LSA	A set of student answers for short essays and relevant model answers	The same dataset for model building and a new dataset	Grade the assignment and evaluate the result	Correlation, adjacent agreement, exact agreement
Shehab et al.	2017	Automated Essay Grading System (AEGS) Framework	Machine learning: neural network	Pre-graded Mansoura University student's essays	Ungraded Mansoura University student's essays	Grading a new essay and evaluate the correlation	The correlation coefficient values for all dataset are larger than 0.7
Jin et al.	2017	Semantic Representations	Machine learning: KNN or SVMRank	Pre-scored essays from ASAP and GoogleNews dataset	Testing set of the dataset	10-fold cross-validation on the training data by random partitioning	Kappa, Pearson correlation coefficient, Spearman correlation coefficient, and normalized root-mean-squared error
Zhao et al.	2017	Memory Networks	Machine learning: neural network	Kaggle ASAP competition	Testing set of the dataset	5-fold cross-validation	QWK: 0.78

Combination of Information Extraction and Machine Learning

Gutierrez et al. (2013) created a hybrid Ontology-Based Information Extraction (OBIE) which combined machine learning and information extraction method in their system to identify the correct and incorrect answer. The performance of hybrid configuration was compared with the pure configuration with two datasets: the real dataset and synthetic dataset. The synthetic dataset was created to explore scalability issues. In the real dataset, the extraction rules have a higher precision than machine learning based extractors. Hybrid configuration with more extraction rule extractors also has higher precision. For the recall, machine learning extractors perform better than extraction rules and so does the hybrid configurations with more machine learning extractors. As a result of the higher result in precision and recall, the hybrid configurations are also having higher F1 result than the pure configurations.

The performance of the approach in the synthetic dataset is also quite similar. For correct information extractions, the precision of pure extraction rules is higher than pure machine learning based extractors. The hybrid method, in which more extraction rules are used, has the trend of higher precision than the pure machine learning based extractors, even higher than the pure extraction rules for 1ML-3ER. The recall of both the pure and mix configuration of more machine learning based extractors outperform the performance of extraction rules in a larger margin. On the other hand, the F1 measure shows the overall average performance from all configurations is similar. The performance of incorrect information extractors is almost identical with correct information extractors, which extraction rules producing better precision and machine learning generating higher recall. The F1 measurement between all methods are alike, but there is more gap between the best and worst performance. The researcher did not explain in depth the reason for this phenomenon.

2.5. Datasets in Previous Work of Automated Grading

Researches in automated grading system have used various datasets. A few of studies used different data for the training and the testing, but most of the studies build the scoring model and evaluate the system using the same dataset because there was a limited size for the datasets.

Based on the exclusion criteria, this study focuses on English dataset only. For essay assignment, most studies used publicly available datasets from Automated Student Assessment Prize competition taken from kaggle.com, while others used essays or answers written by students in schools or university for real examinations. Most of the datasets were scored by human grader(s) before being used in the system so that the performance of the machine grading can be compared with the human grading.

Table 4 Datasets Used in Automated Grading System

Datasets	Works by
Essays from Kaggle Automated Student Assessment Prize competition	Nedungadi et al., 2014; Zupanc & Bosnic, 2014; Dronen et al., 2015; Jin & He, 2015; Phandi et al., 2015; Jin et al., 2017; Mehmood et al., 2017; Zhao et al., 2017
Written answers from real tests	Siddiqi & Harrison, 2008; Wang et al., 2008; Lajis & Azizi, 2010; Ziai et al., 2012; Gutierrez et al., 2013; Omran & Ab Aziz, 2013; Rahimi et al., 2014; Wolska et al., 2014; Perera et al., 2016; Wonowidjojo et al., 2016; Jayashankar & Sridaran, 2017; Shehab et al., 2017
Responses from online system	Latifi et al., 2016; Nakamura et al., 2016
Others (Texas corpus, Chinese Learner English Corpus, TOEFL Corpus)	Bin et al., 2008; Omran & Ab Aziz, 2013; Pribadi et al., 2018; Vajjala, 2018

2.6. The relationship between Machine Learning, Text Mining, and Automated Grading

The most common techniques in automated grading are information extraction and machine learning. Information extraction is one of the text mining technique to identify and extract meaningful information. In automated grading, information extraction uses a series of pattern matching, for example, parse trees, regular expression matching, syntactic pattern matching, and semantic analysis, to find information related to the answer. In machine learning, the answers are pre-processed using text mining concepts, such as tokenization, stop words removal, and stemming, to create the TDM. Then, the scoring model uses the TDM to give scores to the answer.

In other words, the automated grading system of both techniques uses text mining in its implementation. The difference is most of the information extraction techniques tend to involve human grader as the final grader, and to have answer keys for a comparison to the students' answer. The automated grading system acts as the tool to help the grader in determining the score, by showing the important words in the answer or ranking the answer based on its similarity to the answer keys. In contrast, most of the machine learning system relies on the system to mark students answer. The system has an original human score so the performance can be measured by comparing the machine grade with the human grade. Both techniques have similar measurement metric, in terms of accuracy, agreement rate, and correlation. For machine learning technique, there is a common additional measurement metric called kappa value.

2.7. Tools Selection

Most automated grading systems prepare their dataset using pre-processing techniques before it is being processed. Researchers usually use WordNet, Natural Language Toolkit (NLTK), StanfordParser, SVMTool in the pre-processing step. Unfortunately, most studies found in this study did not explain clearly what tools that they used to build the system. Therefore, other studies, that listed what tools are usually used in the data mining process, were explored. There are commercial and open source tools which can be used. Table 5 displays the summary of different data mining tools.

Table 5 Comparison of Tools Used in Automated Grading System

Tool	Features	Advantage(s)	Limitation(s)
Weka	Java-based, open source data mining and machine learning tool.	Suitable for developing new machine learning schemes and can load data file in formats of ARFF, CSV, C4.5, binary. Has a large number of data mining algorithms.	Poor documentation. Lacks many data survey and visualization methods. The support for big data, text mining, and semi-supervised learning is currently limited.
R	Free software programming language and software environment for statistical computing and graphics.	Large-scale statistical library, easier to combine with other statistical calculations. Offers very fast implementations of many machine learning algorithms.	Not a user-friendly environment. Has a steep learning curve. Difficult to learn thoroughly to be productive in data mining.
RapidMiner	Java-based environment for machine learning and data mining processes. Previous version (v. 5 or lower) were open source, while the latest one is not.	Has visually appealing and user-friendly GUI. Has the full facility for model evaluation. More than 1,500 methods for data integration, data transformation, analysis, modeling, and	More suitable for people who are accustomed to working with database files. The support for deep learning methods and some more advanced specific machine learning algorithm is

		visualization.	currently limited.
Orange	Component-based data mining and machine learning open source software, featuring a visual programming front-end and Python bindings.	The easiest tool to learn. Cross-platform GUI. Has better debugger. The shortest script for doing training, cross-validation, algorithms comparison and prediction.	A limited list of the machine learning algorithm. Machine learning is not handled uniformly between the different libraries.
KNIME	Open source data analytics, reporting, and integration platform.	Integrates all analysis modules of the well-known, such as Weka and R-scripts. Requires no installation.	Limited error measurement methods. No wrapper methods nor automatic facility for parameter optimization of machine learning/statistical methods.
Matlab	Commercial data mining tool which contains various packages and toolboxes.	Supports access to a library of popular data handling methods, friendly data visualization, and statistical tools. Good for numerical computations.	The data preparation process is difficult and lengthy because of a lack of toolboxes. No functions for revolving clustering with categorical data. The code in Matlab is more difficult to read and understand.
GATE	Open source software for standard text mining applications, building and annotating corpora, and evaluating the applications.	Provides a variety of tools for text processing and access to various types of linguistic resources. Capability to process text from several different domains and genres.	N/A
LightSIDE	Open source software for text mining.	Gives suggestion to the user to choose what set of attributes is best suited to represent the text. Offers a number of algorithms to perform learning mappings between attributes and the final score.	N/A
Python	Open source programming language which contains a large standard library.	Easy to learn and read. Compatible in any operating system. Offers a wide set of choices in graphics package and toolsets.	When processing large quantities of code to perform numerous actions, the speed and the program's ability to identify and fix semantic errors could be extremely frustrating.

RapidMiner is chosen for this study. It is a complete package for text mining and machine learning. Compared to other tools, RapidMiner is independent of language limitation (comparative study), has an intuitive and interactive GUI, and covers different algorithms. RapidMiner also has an extension for R and python scripting, two other popular programming languages in text mining and machine learning. Therefore, RapidMiner is a powerful tool with its extensions.

CHAPTER 3 – METHOD

The aim of this study is not to grade an answer. Instead, this study analyses the characteristics of an open question and the answer using text mining and machine learning. Those characteristics are beneficial in creating a method in automated grading. Furthermore, the teacher can construct a question that is suitable with the method based on the characteristics. In other words, the automated grading works as a tool to help the teacher in grading task.

Therefore, the method that I propose in this study consists of three elements.

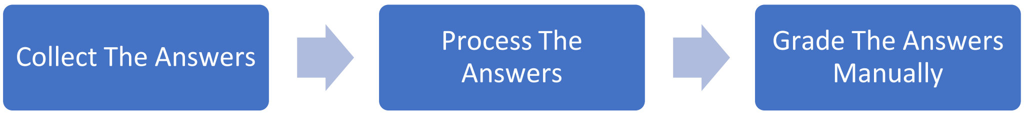


Figure 3 Proposed Method

- In Collect The Answers, the teacher gathers the answers to be graded. The current method assumes that the students write their answers in a digital file, such as in Word, Excel, or PDF. The file is not processed directly, but the answers are separated for each question because each question requires a specific method to evaluate the answer.
- Next, the collected answers are processed using the automated grading. The rest of the thesis explains how this element is implemented for two types of open questions.
- Finally, the teacher grades the answers. The previous phase produces several answers' groups based on the answers' similarity as the results. The teacher can do the grading task for these answers' groups.

Next section describes in detail about the second element by defining some process to assess two types of question.

3.1. Mention a number of examples and explain

In manual grading, the lecturer looks into how many examples are given by the student and the content of the answer. A sufficient number of examples as requested in the question and clear response are an indication of a good answer. The automated system can also use a similar technique as manual grading.

In some cases, a teacher does not give a specific number of examples in the question. This situation leads to answers with numerous numbers of examples. One student could give only one example while the others could explain more than one. It makes the scoring process more demanding because it appears to be unfair to give the same score for both of the students who give the correct answer, but one student only mentions one example, and the another describe 2 or 3 examples.

On that account, besides designing an algorithm of an automated system for this question, this research also examined the effect of the availability of a fixed number being asked in a question to the performance of the automated system.

The question used for this type was taken from e-commerce exam which asks

*"Please give **two examples** of **benefits** and two examples of **limitations** of Blockchain technology."*

For this type of question, there are two main processes. The first process is checking the number of examples and categories in the student answer whether it is sufficient as required in the question or not. Then, check the answer's content by machine learning technique in the second process.

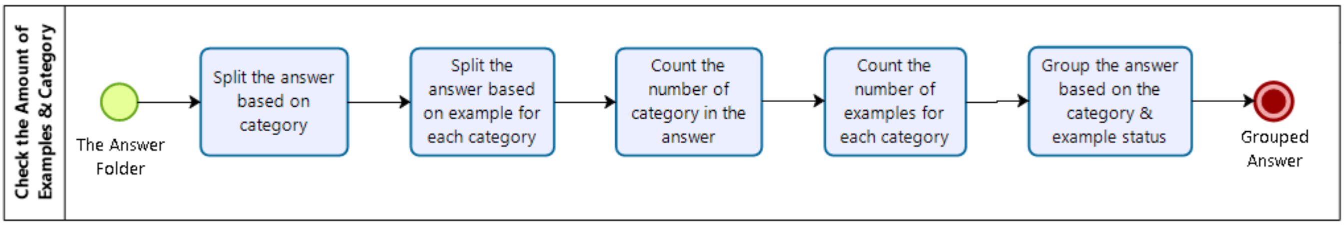

Check the Amount of Examples & Category
The Answer Folder
Split the answer based on category
Split the answer based on example for each category
Count the number of category in the answer
Count the number of examples for each category
Group the answer based on the category & example status
Grouped Answer

Figure 4 First Process: Check the Amount of Examples & Category

Figure 4 represents the flow of the first process. The process is begun by splitting each answer in the received folder based on the category. Keywords and its synonyms used for the category – in this case, for limitations category, words like downsides, disadvantages, or drawback – determine the result of the split. Then, count the result of category splitting. After that, divide the category based on the examples mentioned in the answer using numbering, bullets, or new line, to calculate the number of examples mentioned in the answer. Finally, combine the result of category and examples separation and use it to group the answers.

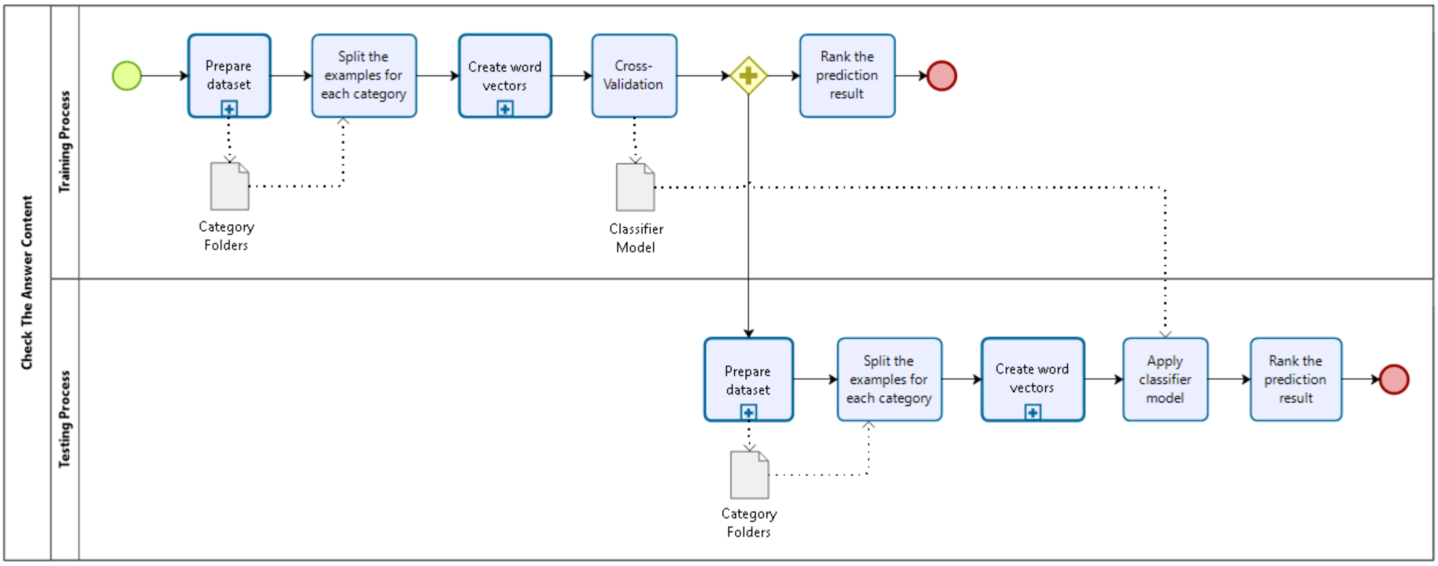


Figure 5 Second Process: Check The Answer Content

The second process is checking the content which contains training and testing process as illustrated in Figure 5. Preparing dataset for the training process is the first step in this process, which is depicted in Figure 6. Since there are two categories (benefits and limitations) in this case, the result of preparing the dataset is two different folders consist of all answers in each category. These folders and the files within are used for further processing.

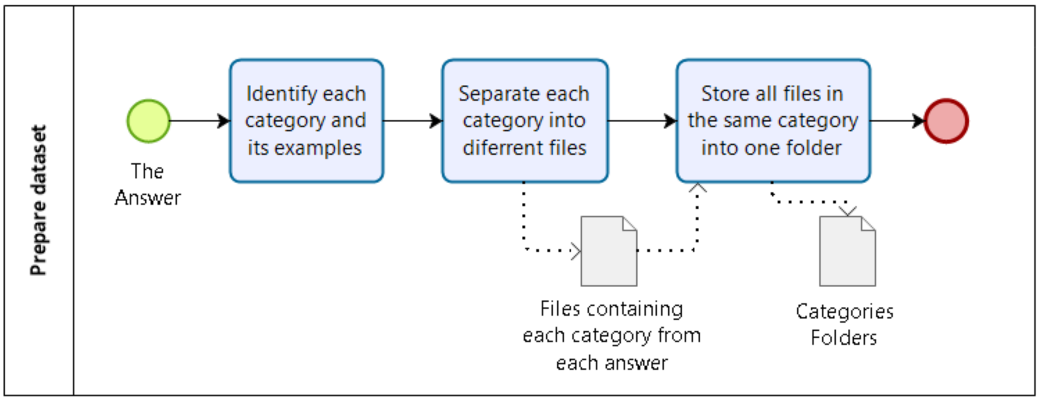

Prepare dataset
The Answer
Identify each category and its examples
Separate each category into diferrent files
Store all files in the same category into one folder
Files containing each category from each answer
Categories Folders

Figure 6 Second Process: Inside Prepare Dataset Process

Next, divide each category file into several examples. Use the examples to be processed by text mining techniques to create the word vectors as shown in Figure 7. After that, apply cross-validation to build the classifier model using the training data and evaluate the performance. After the process in training process complete, repeat the same technique for the evaluation process. The difference is, instead of executing the cross-validation technique, apply the classifier model to the evaluation data and examine the result.

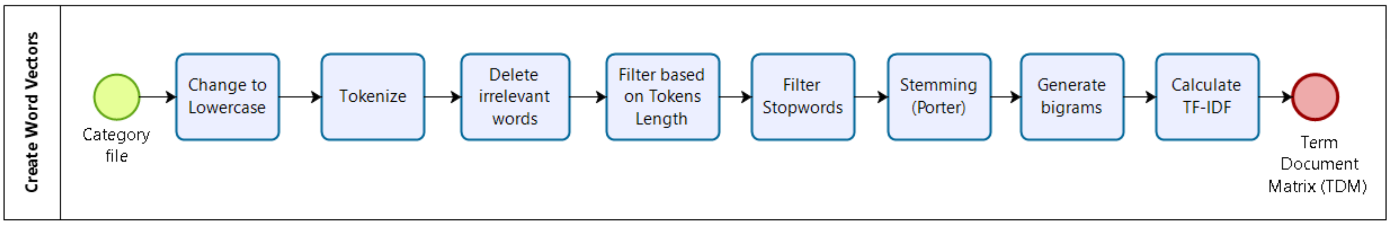


Figure 7 Second Process: Inside Create Word Vectors Process

The process of checking the answer content does not give a score to the answers, but only examines the variety of students' answers in each category. The higher accuracy value means higher possibility that the students' answers are similar to each other, hence the model can classify the answers well. After the classification, the answer is ranked based on the confidence value. The higher confidence value of the real class implies a higher likelihood of similarity with the others. Therefore, an answer which has the correct prediction results and high confidence value indicate a higher similarity. Human graders have options to grade in two ways: based on the sufficient number of examples or the ranking. Figure 8 and Figure 9 describes the ranking process.

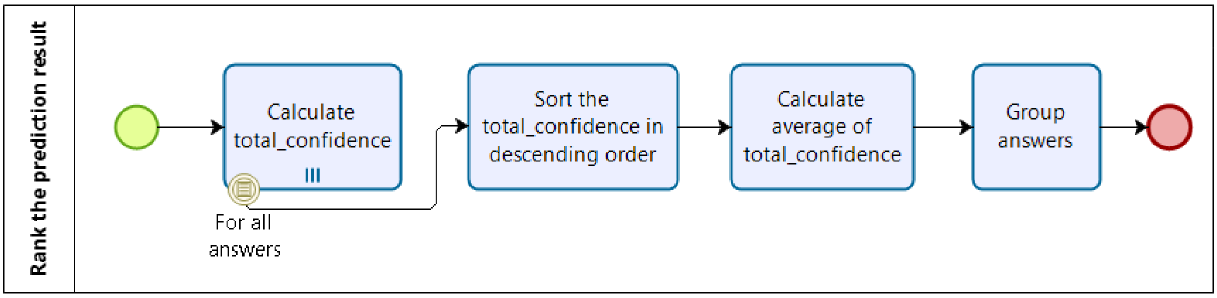

Rank the prediction result
Calculate total_confidence
For all answers
Sort the total_confidence in descending order
Calculate average of total_confidence
Group answers

Figure 8 Ranking Process

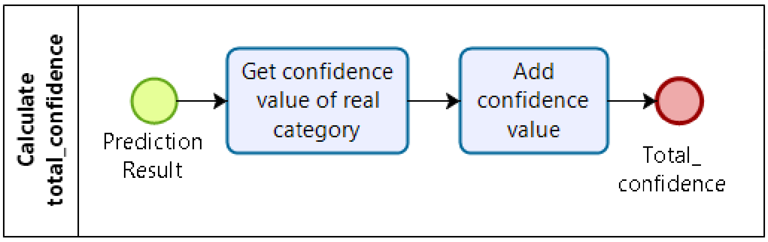

Calculate total_confidence
Prediction Result
Get confidence value of real category
Add confidence value
Total_ confidence

Figure 9 Ranking Process: Inside Calculate total_confidence Loop

3.2. Ask opinion from the students

A question that asks the opinion from the students about a concept receives various answers because students have their different thoughts. Even a question that might contain general and particular concept(s) could be responded in different style. Therefore, it is complicated to grade this kind of question, either manually or automatically.

Previous works in automated grading which are similar to this set of question are automated essay grading systems (Bin, Jun, Jian-Min, & Qiao-Ming, 2008; Perera, Perera, & Weerasinghe, 2016; Phandi, Chai, & Ng, 2015; Shehab et al., 2017; Wonowidjojo, Hartono, Frendy, Suhartono, & Asmani, 2016). The system marks the essay with classification techniques by using features of the essay, such as the grammar, word vectors, or word count. Most of the systems relied on supervised machine learning techniques and more than 50 sample data. However, not all courses in the higher-education system have students more than 50 people. These methods are not suitable for small classes.

Therefore, this study explored another approach through clustering to assess this question. Clustering does not need training data during the learning process. It groups the data based on a particular characteristic(s) and similarities between the data, while different data is clustered to another group. The determinant factors used in this research are sentiment analysis polarity, length of the answer, and the TF-IDF word vectors. Figure 10 below shows the process for this question type.

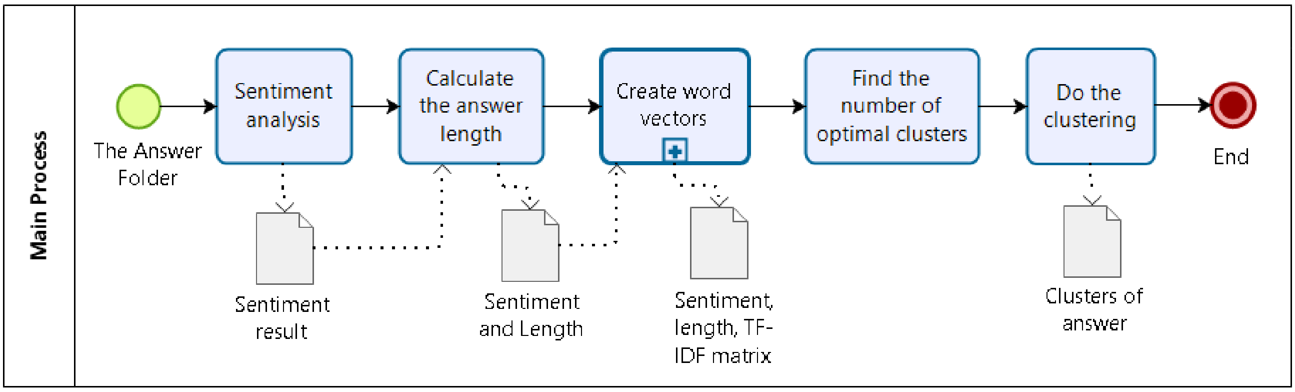


Figure 10 Main Process of Question Type 2

The process receives the folder and answers files inside it. Next, perform sentiment analysis to gather information about student's opinion on the topic asked in the question. Sentiment analysis is one example of text mining application to determine the feeling expressed in a text. The feeling is divided into positive, negative, and neutral (Justicia De La Torre, Sánchez, Blanco, & Martín-Bautista, 2018; Kent, 2014). There is also polarity confidence as the result of the analysis. The opinion and the polarity confidence value are retained for the clustering. After the sentiment analysis, calculate the length of each answer to be included in the clustering process.

Then, create the word vectors. In this phase, the answers are pre-processed using NLP techniques as displayed in Figure 11. The result of this subprocess is a TDM based on TF-IDF.

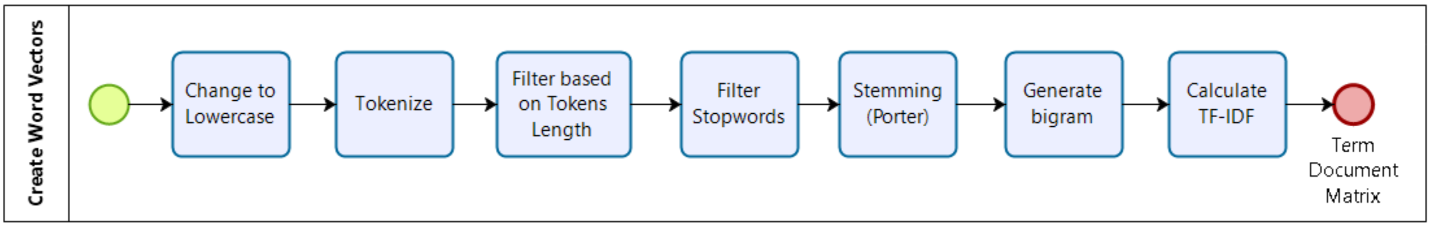


Figure 11 Inside Create Word Vectors Process

- “Change to lowercase” transforms all letters into lowercase because the TF-IDF calculation is case-sensitive. To reduce the number of words, changing all words into lowercase is needed.
- Tokenize splits the text into a list of words because the delimiter specified is non-letter characters. Therefore, all numbers, symbols, and whitespaces are removed.
- The tokens of words are filtered based on its length. In this study, the length of the token to be considered as meaningful information for the clustering is three characters or more.
- Stop words are also removed to preserve only relevant words into the calculation.
- Some words have the same root. Therefore, stemming is applied to reduce the number of word lists. This study used the Porter algorithm to do the stemming.
- Because the combination of two words might contain useful information than one word only, unigram and bigram are generated to be included in the calculation.

Determining the number of optimal clusters is difficult, especially when there is no exact number specified beforehand. Davies-Bouldin (DB) index becomes an option to find out the optimal clusters. DB index is a function of the ratio of the sum of within-cluster scatter to between-cluster separation (Bandyopadhyay & Maulik, 2001). The smaller value of DB index is preferred since it indicates more compact and well-separated clusters (Visvanathan, Srinivas, Lushington, & Smith, 2009). Therefore, the *k* value with the smallest DB index value is chosen as the optimal clusters.

Finally, the clustering is executed, and human grader analyses the result. This study used X-Means algorithm to cluster the answers. X-means algorithm is found by Dan Pelleg and Andre Moore in 2000 to overcome limitations in the k-means algorithm (Pelleg & Moore, 2000). The algorithm is faster than k-means and computes the number of clusters dynamically using the lower and upper bound supplied by the user (Kumar & Wasan, 2010). The algorithm searches the space of cluster locations and a number of clusters efficiently by optimizing Bayesian Information Criterion (Alickovic & Babic) or The Akaike Information Criterion (AIC) measure (Kumar & Wasan, 2010).

This study selects Silhouette index as the measurement metric. Silhouette index is a validity index to examine the quality of the clusters through cohesion and separation (Errecalde, Cagnina, & Rosso, 2015; Pérez-Delgado, Escuadra, & Antón, 2010). Cohesion indicates how similar are the objects within the cluster, while separation signifies how different a cluster to each other is (Errecalde et al., 2015). A value close to -1 indicates the object is clustered into the wrong cluster; if the value is near to 0, then the object is in the border of the cluster, and it is not clear whether it really belongs to its cluster or should be placed into its neighbor; and value close to 1 or higher denotes a good cluster (Shanie, Suprijadi, & Zulhanif, 2017; Visvanathan et al., 2009).

CHAPTER 4 – EXPERIMENT DESIGN AND THE IMPLEMENTATION IN RAPIDMINER

This research uses RapidMiner as the tools for implementing the methods explained in the previous chapter. There are various products of RapidMiner, and this study uses RapidMiner Studio version 9.0.003 with student license. RapidMiner contains around 400 built-in operators for data mining and machine learning. It also supports more extensions, such as text processing, AYLIEN Text Analytics, and web mining, from the marketplace to enrich the analysis process. Furthermore, RapidMiner can read various types of file, from a text file, Excel, PDF, XML, and HTML. For this study, Text Processing and Text Analysis by AYLIEN extensions are downloaded from the Marketplace.

People who only have basic knowledge about the database, data mining, and machine learning can use RapidMiner easily because users do not have to code their own program in RapidMiner. They just need to drag-and-drop the data and the operators to create the process flow. Because of its simplicity, RapidMiner features were chosen to assist the process in creating the method for automated grading in this research.

There are several terms in this chapter related to RapidMiner application. The list below explains these terms:

- Operator: a function in RapidMiner to execute a particular task. For example, Retrieve operator accesses stored information in the Repository and load them into a Process. An operator is represented in a single box. An operator can be connected to the others to create a process.
- Parameter: an operator can have none until more than one parameter. For example, Retrieve operator has one parameter for the repository entry. Retrieve operator get and load the object in the repository specified into the Process. Transpose operator does not have any parameter because it just switches the row into column and vice versa.
- Example: similar to a row in a database. One example set is defined by its attributes and can be processed (compared, aggregated, or filtered) with other examples within an Example Set as long as both of them have the same attributes.
- Attribute: similar to column in a database.
- Example Set: a collection of multiple examples. It is similar to a table in a database.
- Collection: a group of multiple Example Set.
- Subprocess: a process within a process. It consists of an operator or a series of connected operators doing an action. Several operators are a subprocess, such as Cut Documents, Loop Files, and Branch. Subprocess operator or an operator that is a subprocess is represented in a double box.
- Aggregate operator: performs a similar function as the aggregation function in SQL. The operator executes aggregation function, such as average, concatenation, and count, by the aggregate attribute and the results can be grouped by selected attribute(s).

4.1. Dataset

This study works on datasets from actual student answers in the final exam from two courses in Business and IT program: e-commerce and Business Case Development for IT (BCD4IT). There are 30 student answers from e-commerce and 47 answers from BCD4IT. The dataset was given in Word documents. Before being processed in RapidMiner, the corresponding answer was copied to a text file.

From the dataset, there are three types of question taken for the experiment. According to Bloom, teachers have a tendency to ask a question in knowledge level for 80% to 90% of the time (Ahmad et al., 2011; Clay, 2001). This is not a good example because a good exam consists of different levels of understanding of Bloom's taxonomy to facilitate the students developing critical thinking ability (Ahmad et al., 2011). For that reason, these questions are chosen because these types of questions are typical in open questions exams and include not only knowledge level, but also other higher levels.

a. A question asks students to mention a number of examples and explain it. This type of question assesses knowledge, comprehension, and application level on Bloom's taxonomy. The lecturer not only can use this question to general knowledge, but also can apply it into a case study question. For example, instead of only asking the students about mentioning the benefits of using blockchain in general, the lecturer can also ask the student to write down the benefits in the specific business area, like in music industry, and describe the reason why blockchain is counted as a benefit.
b. A question asks the opinion from the students. This question can provide a proposition or a concept, then ask what the students think about it and why. Students can agree, disagree, or neither agree nor disagree. Through this question, the lecturer can examine how well the student understands a concept that is being asked. In Bloom's taxonomy, this question might involve all six levels depending on the purpose of the question. An example of this type is "Would the effectiveness of a recommender system benefit from a well-designed and well documented pluggable business/IT architecture? Motivate your answer".

4.2. Whole Process

During this study, all types of question were processed in NLP technique, but in different approaches, because their characteristics are different. Figure 12 and Figure 13 depict the overall process implemented in this study.

1. The answers were anonymized and given in the form of Word files. The files were named as "StudentXX.doc" with XX denotes a number starts from 1.
2. Because RapidMiner cannot read files in any Word extensions, the answer should be written in text (.txt), PDF, or Excel file, and this study selects text files. Therefore, the answers were copied and pasted into separate text files and stored in one folder for each question. The methods were applied to the folder and all files inside it (Loop Process the answers).
3. There are two types of answers researched in this study. The first answer is mentioning a number of examples and explaining them. The second one is about student's perspective on a given statement. Each answer is evaluated in different techniques.
4. Each method generates a result. The researcher can analyze the result and take conclusion for each method.

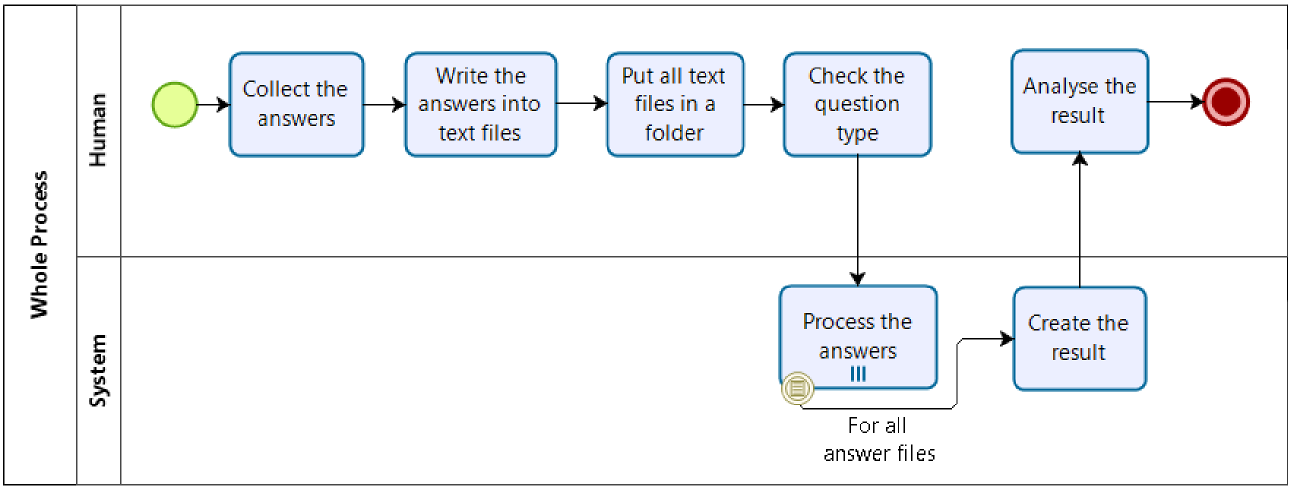


Figure 12 Whole Process

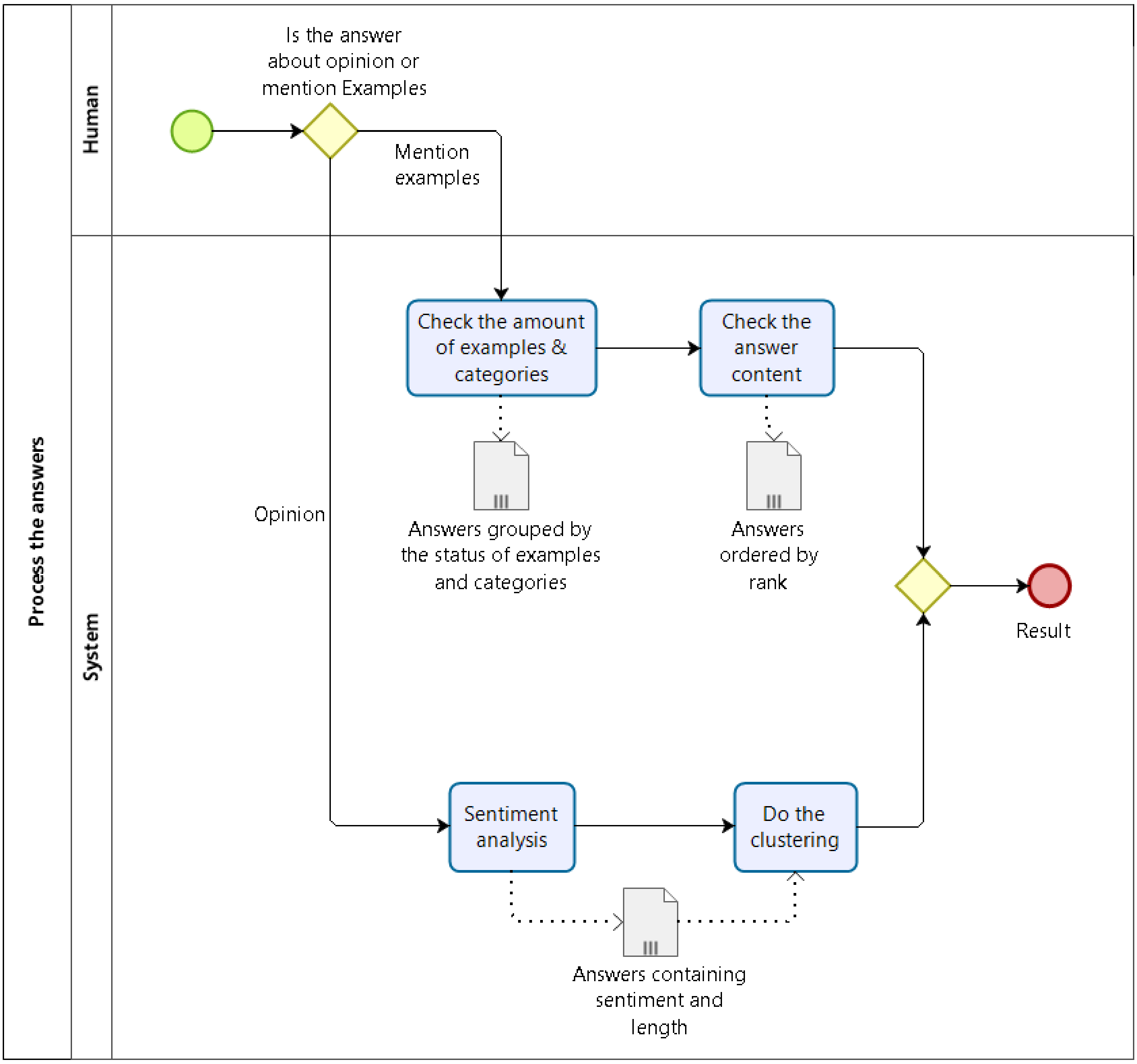

Process the answers
Human
System
Is the answer about opinion or mention Examples
Mention examples
Opinion
Check the amount of examples & categories
Check the answer content
Answers grouped by the status of examples and categories
Answers ordered by rank
Result
Sentiment analysis
Do the clustering
Answers containing sentiment and length

Figure 13 Whole Process: Inside Process The Answers Loop

4.3. Mention a number of examples and explain

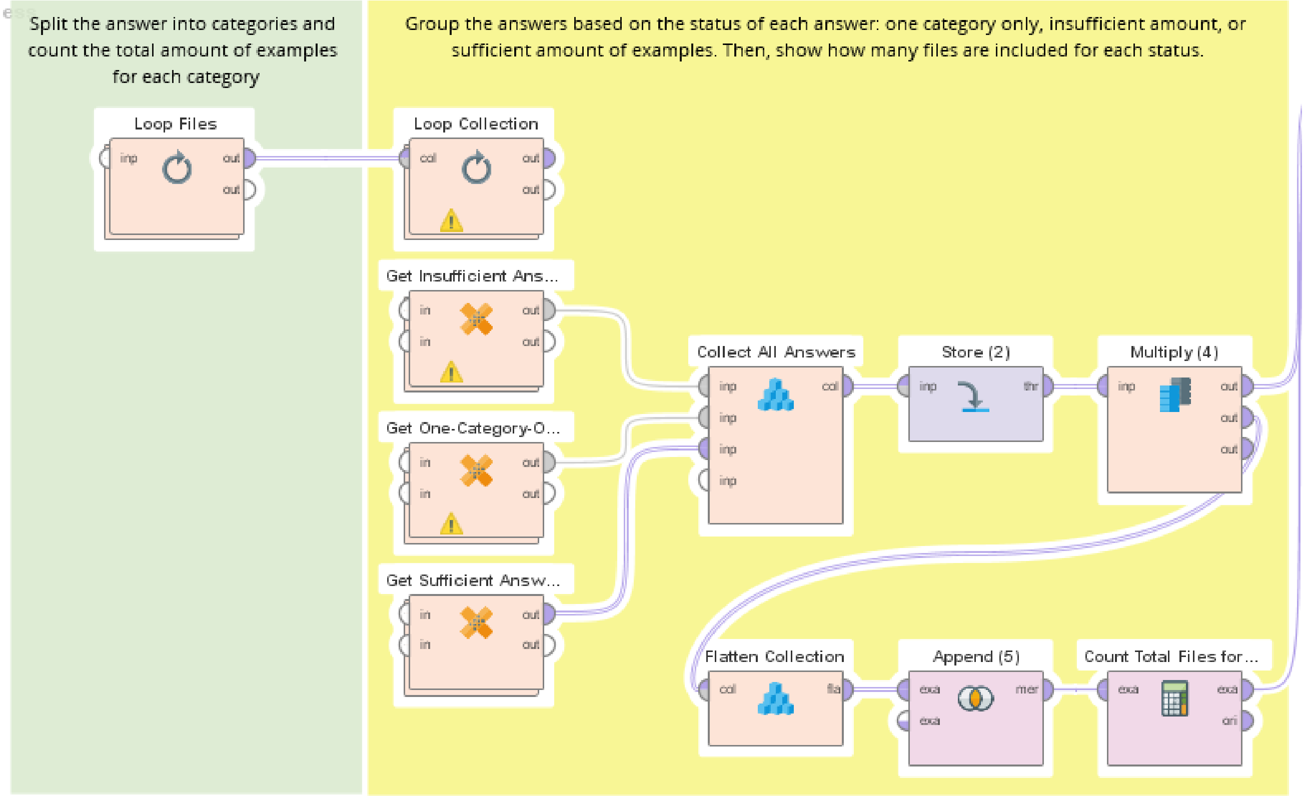
In this type, there are two main processes implemented. First, the method to count the number of examples given by the students: is it sufficient as required in the question or not? Second, the method to assess the content of the answer using machine learning. For this type, selected questions are from both exams. A question from BCD4IT exam illustrates the implementation in this chapter.

Question 2:

> a. What are in your opinion <u>three</u> **strong points** of this business case?
> b. What are in your opinion <u>three</u> **weak/improvement points** of this business case?

4.3.1. Split the category and count the number of examples for each category

The first method is divided into two steps, which are splitting the answer into the category indicates in the question and counting the number of examples available in the answer for each category. In this case, strong and weak points are the categories and three is the total examples being asked. Figure 14 describes how the first method is implemented in RapidMiner.

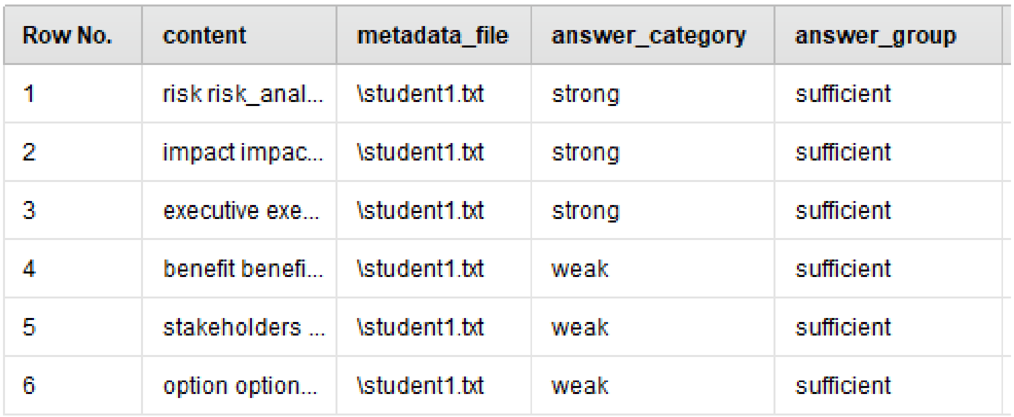


Split the answer into categories and count the total amount of examples for each category

Group the answers based on the status of each answer: one category only, insufficient amount, or sufficient amount of examples. Then, show how many files are included for each status.

Figure 14 Whole Implementation of First Process

The first method is implemented in the Loop Files operator. Loop Files operator requires a directory which is the path of the folder of the answer files. The results of the Loop Files operator are a collection of answers. The answers are already divided into strong and weak points, and an indication of sufficient or insufficient number of categories and examples. Figure 15 shows one result of the Loop Files operator.

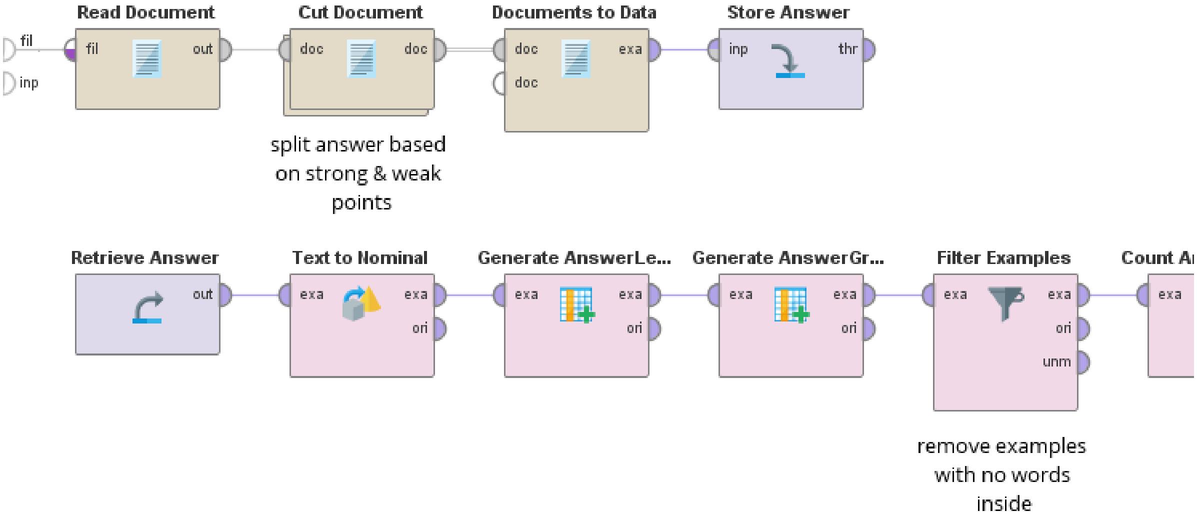


Row No.	content	metadata_file	answer_category	answer_group
1	risk risk_anal...	\student1.txt	strong	sufficient
2	impact impac...	\student1.txt	strong	sufficient
3	executive exe...	\student1.txt	strong	sufficient
4	benefit benefi...	\student1.txt	weak	sufficient
5	stakeholders ...	\student1.txt	weak	sufficient
6	option option...	\student1.txt	weak	sufficient

Figure 15 An Example of Loop Files Operator Results

Figure 16 and Figure 17 demonstrate the tasks inside the Loop Files operator.

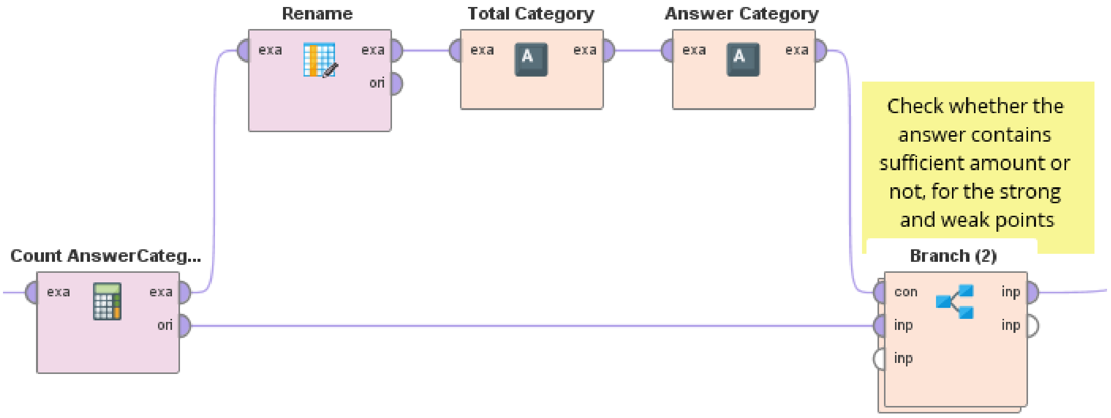
1. The process is begun with reading the file in the directory. Then, cut the file into segments to split it into strong and weak categories, and to get meaningful words in the answer. The categories might be different for another question and should be specified manually using the regular expression in this operator. The results of Cut Document, which consist of the answer category and simplified answer, are transformed into Examples Set and stored (operator Store Answer).
2. The Answer is retrieved, then calculate the length to remove empty and insignificant Example Set. Because the previous process sometimes produces one Example which is not included as strong or weak points, it has to be removed to avoid miscalculation.
3. Count AnswerCategory operator is an Aggregate operator to count how many categories exist in the answer. It only counts distinct value, and the result of aggregation is used as condition expression in the Branch operator next.



split answer based on strong & weak points

remove examples with no words inside

Figure 16 Process Inside Loop Files Operator (1)



Check whether the answer contains sufficient amount or not, for the strong and weak points

Figure 17 Process Inside Loop Files Operator (2)

Inside the Branch operator, there are several conditions to state a sufficient number of examples written in the answer or not. Figure 18 below displays the process inside the Branch operator. Detail implementation in RapidMiner can be seen in Appendix A.1. The result of Branch operator is the status of each category:

- One category only is when only strong or weak point found in the answer
- Insufficient appears when one or both of the category have less than three examples
- Sufficient is when one or both of the category contains three or more examples

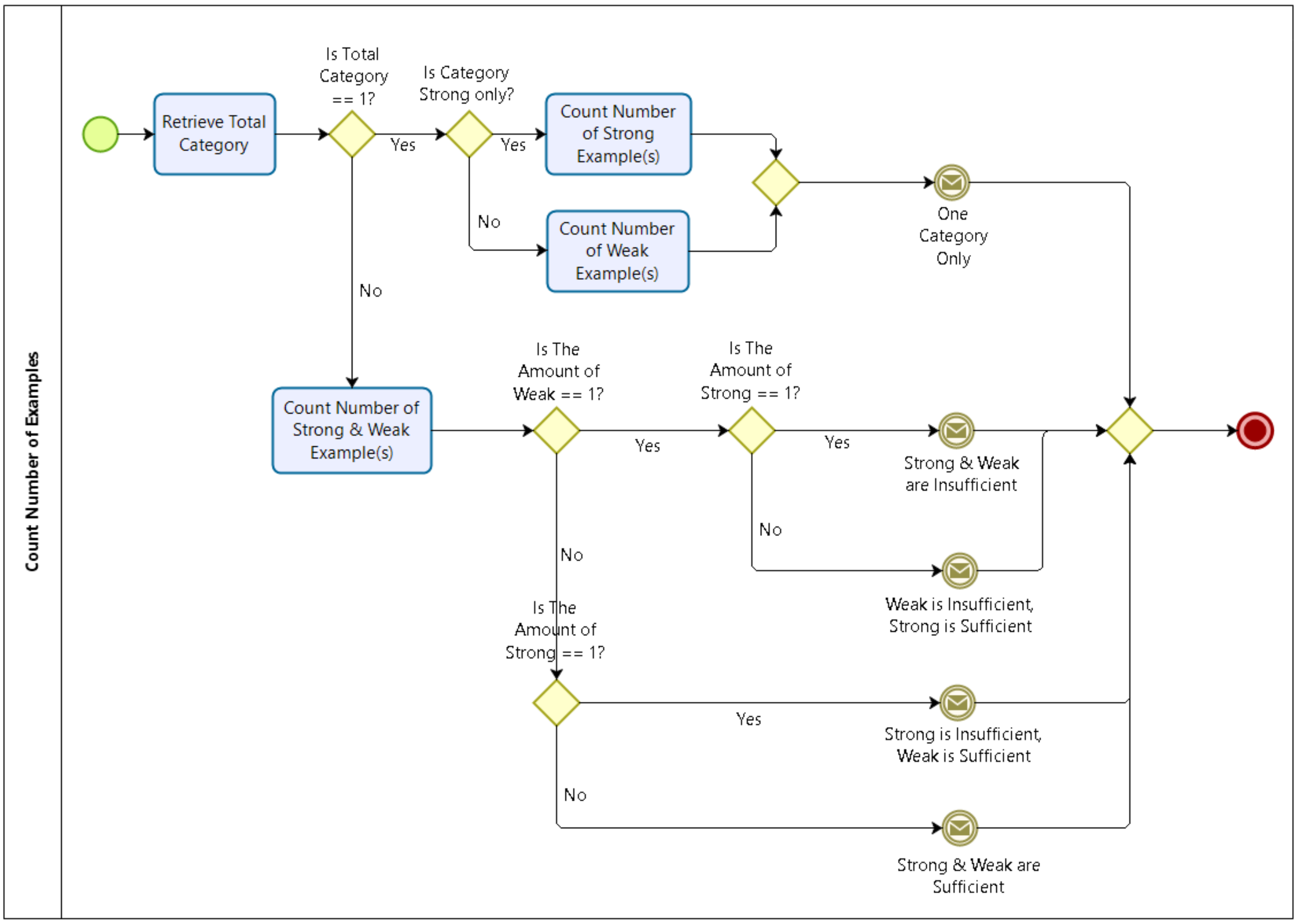
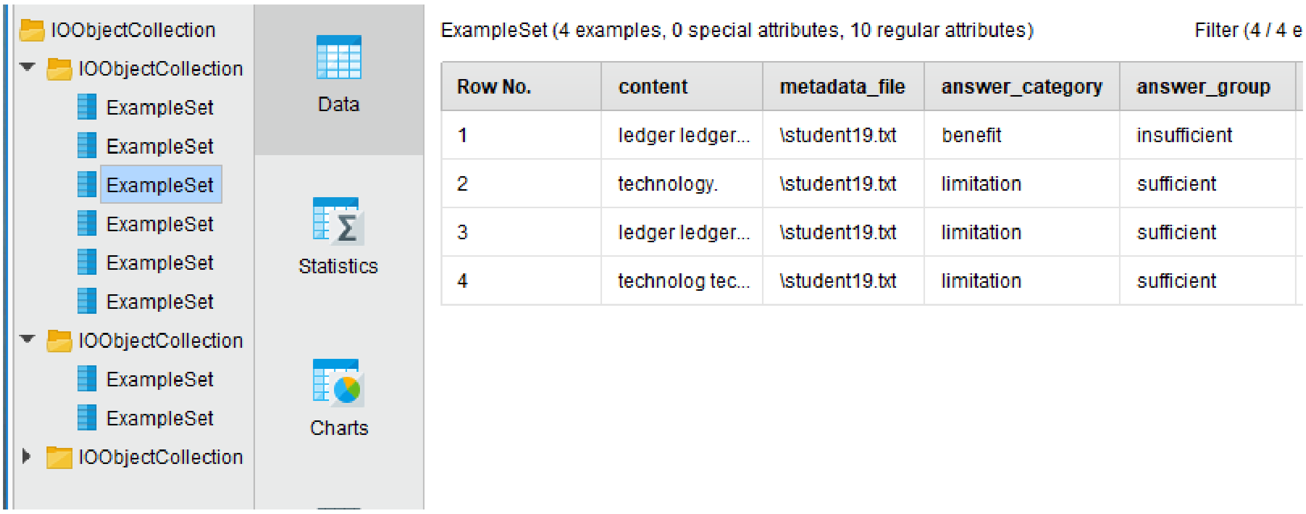

Count Number of Examples
Retrieve Total Category
Is Total Category == 1?
Yes
Is Category Strong only?
Yes
Count Number of Strong Example(s)
No
Count Number of Weak Example(s)
One Category Only
No
Count Number of Strong & Weak Example(s)
Is The Amount of Weak == 1?
Yes
Is The Amount of Strong == 1?
Yes
Strong & Weak are Insufficient
No
Weak is Insufficient, Strong is Sufficient
No
Is The Amount of Strong == 1?
Yes
Strong is Insufficient, Weak is Sufficient
No
Strong & Weak are Sufficient

Figure 18 Process of Counting Number of Categories and Examples

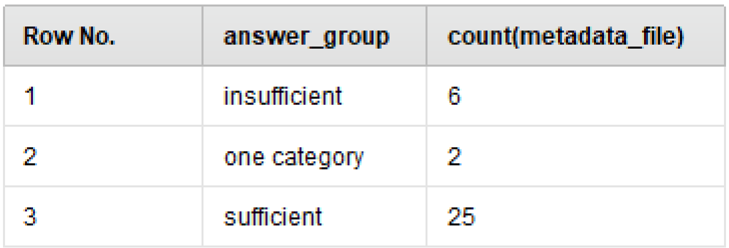
To make it easier for the human grader(s) if they want to check the answer, the answers are grouped based on the status. The grouping task is executed in yellow area in Figure 14. Operators from Loop Collection until Collect All Answers perform the grouping, while operators Flatten Collection until Count Total Files for Each Status generate the total files for each status. Figure 19 shows the result of grouping task, and Figure 20 summarizes the results of the grouping.



ExampleSet (4 examples, 0 special attributes, 10 regular attributes) Filter (4 / 4 e

Row No.	content	metadata_file	answer_category	answer_group
1	ledger ledger...	\student19.txt	benefit	insufficient
2	technology.	\student19.txt	limitation	sufficient
3	ledger ledger...	\student19.txt	limitation	sufficient
4	technolog tec...	\student19.txt	limitation	sufficient

Figure 19 The Results of Answers Grouping



Row No.	answer_group	count(metadata_file)
1	insufficient	6
2	one category	2
3	sufficient	25

Figure 20 The Summary of Answers Grouping Results

4.3.2. Check the answer content

Classification technique is implemented to assess the content of the answer. Since classification is supervised learning, the dataset was separated into training sets and evaluation sets with portion 70% to 30%, respectively. Figure 21 depicts the training process.

Process Documents from Files operator processes text files within the directory and creates word vectors from the files for each folder. Unlike the first method, the answers for this method are already split between the categories. In other words, the inputs for the operator are two folders containing strong and weak points from the students. Those folders are considered as a label for each file in the related folder.

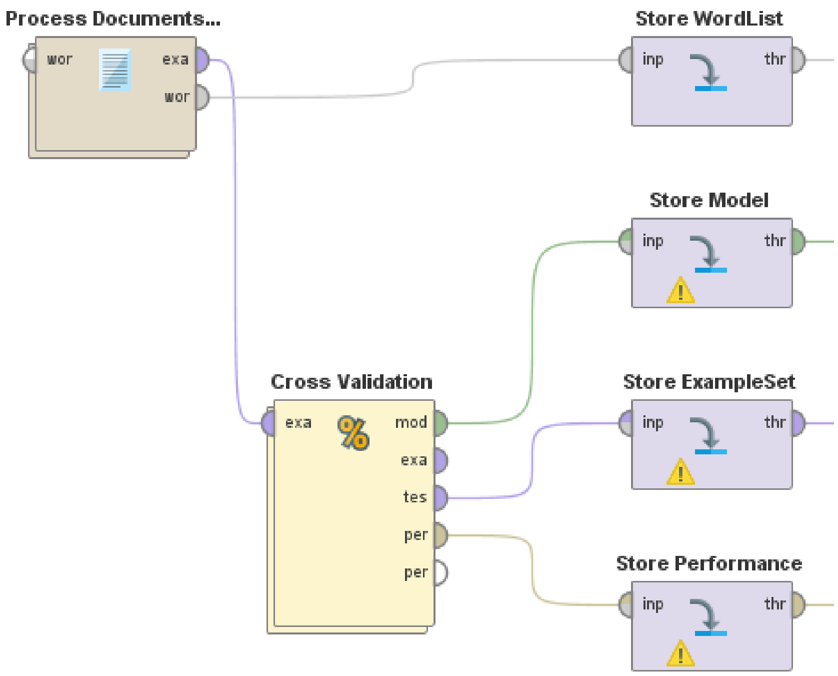
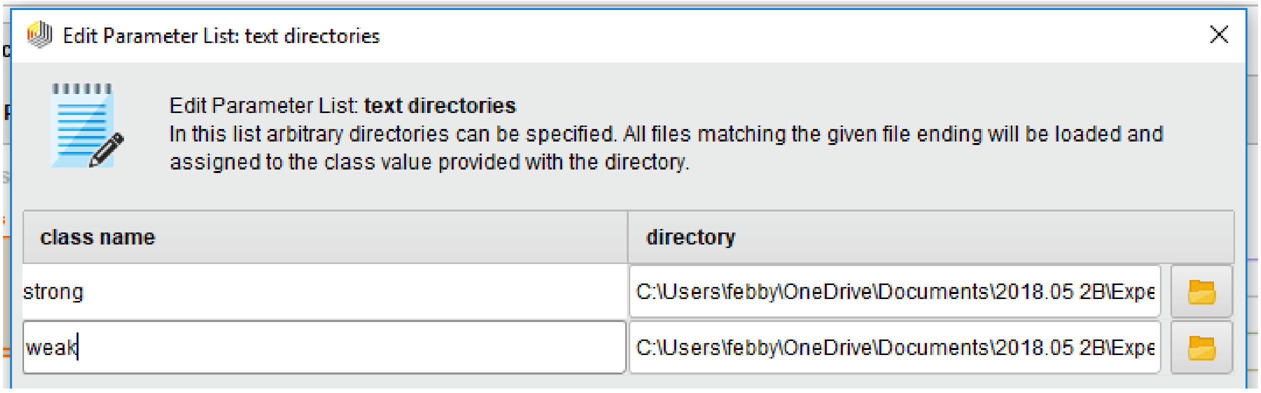


Figure 21 Check The Answer Content: Training Process



Edit Parameter List: text directories

Edit Parameter List: **text directories**
In this list arbitrary directories can be specified. All files matching the given file ending will be loaded and assigned to the class value provided with the directory.

class name	directory
strong	C:\Users\febby\OneDrive\Documents\2018.05 2B\Expe
weak	C:\Users\febby\OneDrive\Documents\2018.05 2B\Expe

Figure 22 Training Process: Specify the Directory in Process Document from Files Operator

Figure 23 and Figure 24 shows the processes inside the operator Process Documents from Files. This is a process to transform the answers into a TF-IDF matrix. This study uses TF-IDF as the word vectors because it is widely used in the previous works and it can control more common words achieve a higher score than less common words, which could be important (Bafna et al., 2016; Vijayarani et al., 2015). Then, send the TF-IDF matrix to Cross Validation operator to train the model. Store the model and the Word Lists for future use in the evaluation phase.

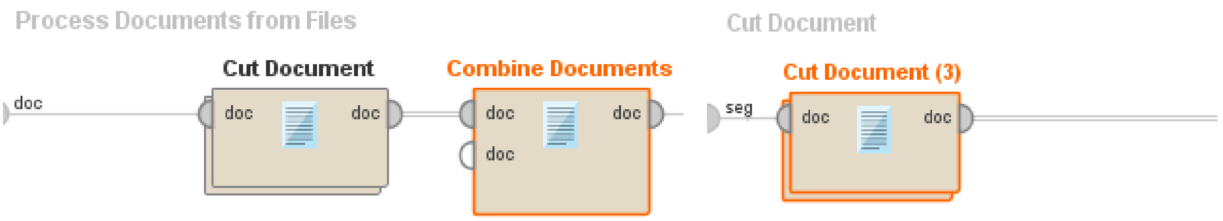


Figure 23 Training Process: Inside Process Documents From Files Operator

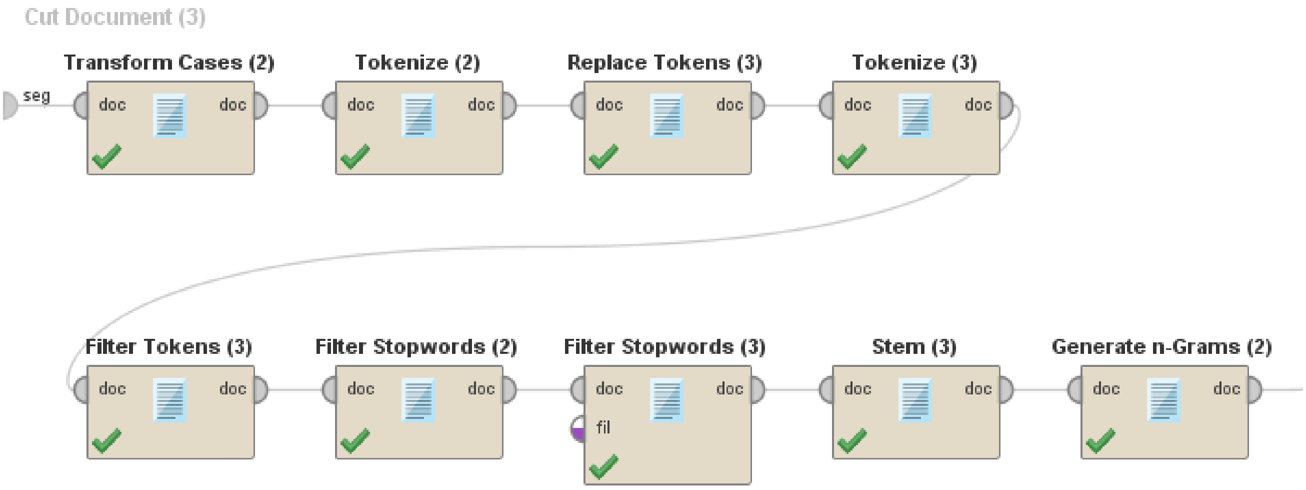


Figure 24 Training Process: Answers Pre-processing Inside Cut Document (3) Operator

Figure 25 demonstrates the learning process in Cross Validation (CV) operator. The experiment tested three common machine learning techniques for textual data which are Naïve Bayes, KNN, and SVM.

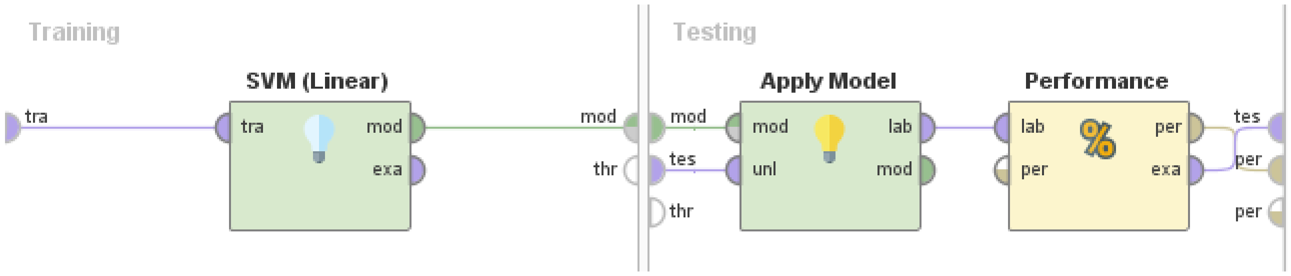


Figure 25 Training Process: Subprocess Inside Cross Validation Operator

- The number of folds in the CV is 5.
- The parameter(s) for each algorithm:
  - Naïve Bayes operator: no changes. Just make sure the laplace correction box is ticked.
  - k-NN operator: set *k* value with 5, NumericalMeasures as measure types, and CosineSimilarity for the numerical measure.
  - No changes for the parameters, except no scaling, the C value is 1.0, and convergence epsilon is 0.01. The changes were made because after several experiments, these values produce the optimal performance.
- Important performance values to be checked in Performance operator are accuracy and correlation.

After the learning process, the next step is applying the SVM model to the new data which is reserved from the beginning of the process as evaluation sets. The evaluation process also creates a TF-IDF matrix from the evaluation sets and word lists from the learning process. After creating the TF-IDF matrix, apply the SVM model to the evaluation sets and examine the performance.

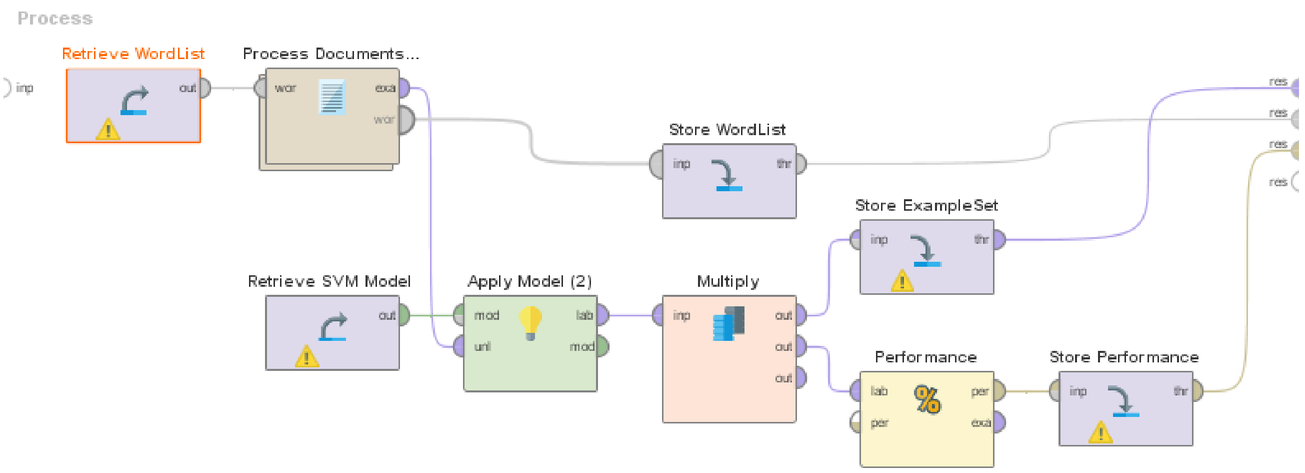


Figure 26 Check The Answer Content: Evaluation Process

Finally, the last process is to rank the answer based on the confidence value. From the classification process, there is a confidence value that expresses the level of model's confidence to classify an example into a class. The higher value determines more certainty the model classify an example into the correct class.

The ranking mechanism is described as follow. The selected confidence value is always the confidence value of the real class. If a strong example is classified correctly as strong, the confidence value of strong prediction is taken. However, if a strong is classified as weak, the confidence value of strong prediction is chosen. The confidence values of strong and weak examples are added, and the result is sorted in descending order to determine the answer's rank. The higher value is assumed to have a good score because the model is more confident in classifying into the correct class. The whole process of ranking can be seen in Appendix A.2.

4.4. Ask opinion from the students

For this question, there are two steps implemented. First, the sentiment of the answer is analyzed. Next, the answers were prepared for clustering. Answers from question 1 in 2018 BCD4IT exam are selected.

Give your view on the proposition below:

The rise of Internet of Things (IoT) and Big Data will significantly impact the field of Business Case Development.

Illustrate your position with an example.

4.4.1. Sentiment Analysis

The Text Analysis by AYLIEN extension from the marketplace is required to do the sentiment analysis. AYLIEN extension is free for the public with a limited daily call to the API. When the limit is reached, the user has to wait for the next day to be able to use it again. Since calling to AYLIEN API requires the API key, a connection must be established. The setup for the connection can be seen in the following link https://developer.aylien.com/signup?source=rapidminer.

Sentiment analysis using AYLIEN is easy. Use Loop Files (Sentiment Analysis) operator to apply the analysis to all answer files. Inside the Loop Files operator, read the answer and calculate its length first. Tokenize operator breaks down the document into a sequence of word tokens, then Extract Token Number extracts the number of tokens. Since Tokenize operator has split the document into tokens, the tokens need to be collected again into one document to do the sentiment analysis by using the Combine Document operator. In Analyze Sentiment operator, select the connection that has been previously configured and document as sentiment mode. The details of the process are shown in Figure 27 and Figure 28.

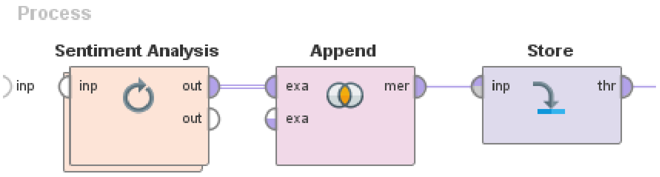


Figure 27 Sentiment Analysis Process

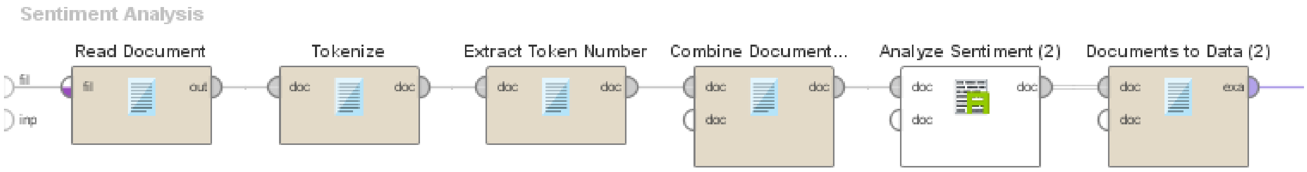


Figure 28 Process Inside Sentiment Analysis Operator

4.4.2. Clustering

The following step after the sentiment analysis is pre-processing the answer and creating the TF-IDF matrix. Figure 29 and Figure 30 illustrate the process. First, retrieve the result of sentiment analysis and counting answer length. Second, map the polarity into numeric value because the clustering algorithm does not consider nominal value into the calculation. The mapping is as follows: negative is mapped into -1, neutral into 0, and positive into 1. Next, use Process Documents from Data operator to pre-processes the answers and create the TF-IDF matrix.

The filename, which is a nominal attribute, is not needed in the clustering, yet it is important information about the examples. Therefore, generate an ID (an integer value and a special attribute) and store the ID and the filename in Store (3) operator to avoid missing the filename. Then, store all numerical features, the TF-IDF matrix, answer length, polarity, and polarity confidence in Store (3) operator as one object for the clustering.

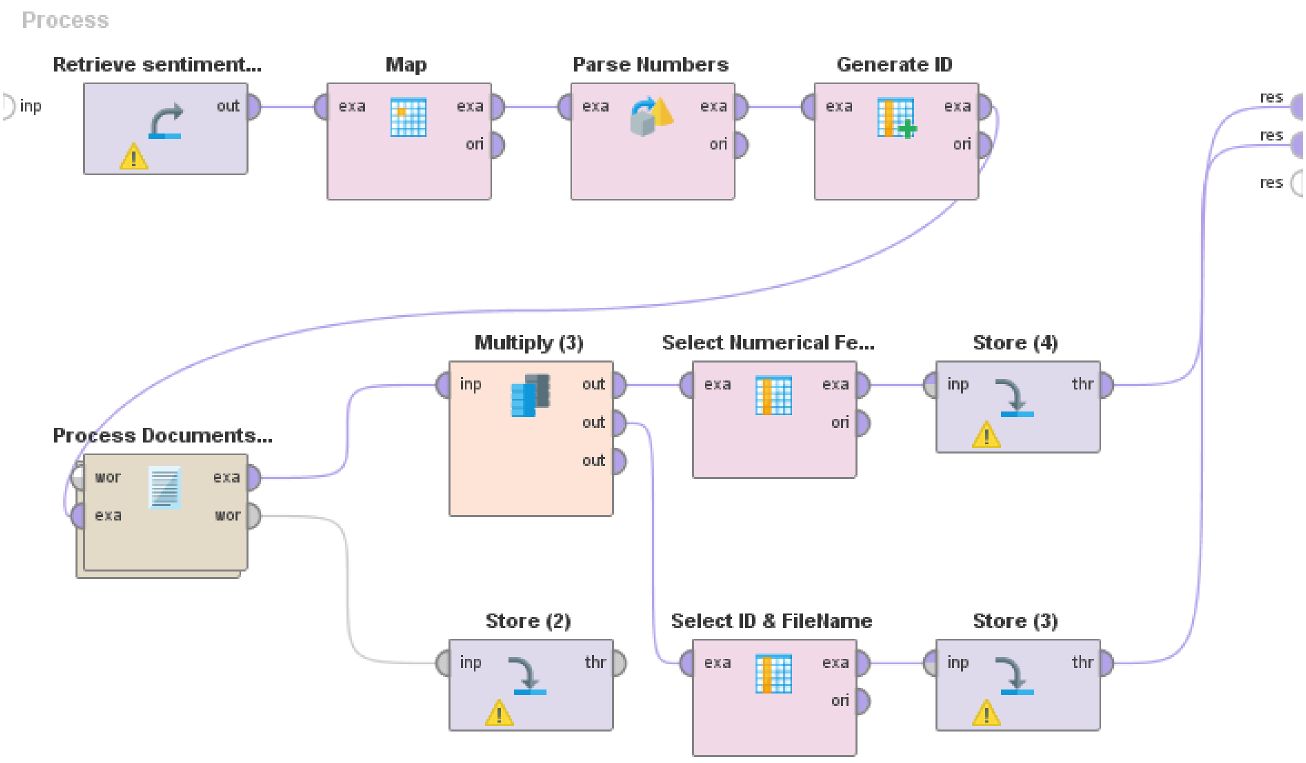


Figure 29 The Clustering Process

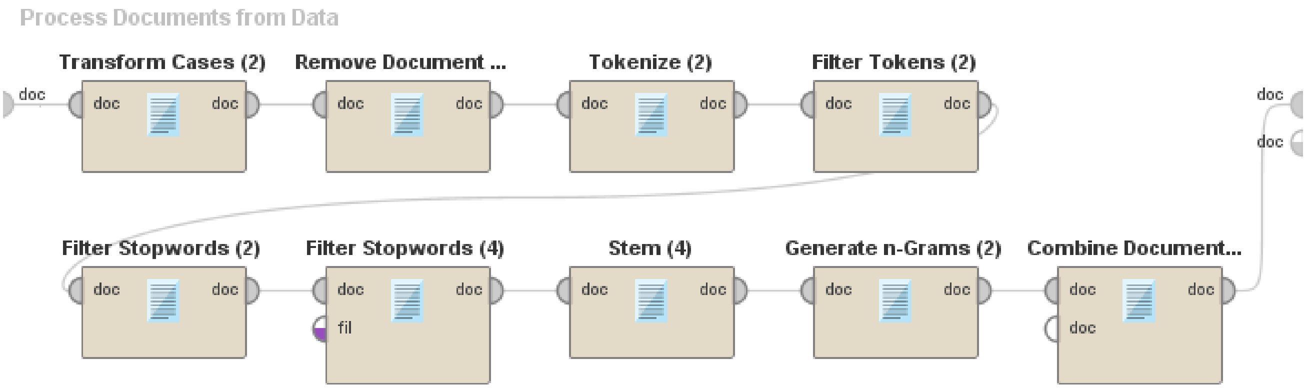


Figure 30 Answers Pre-processing Inside Process Documents from Data Operator

Dimensionality reduction could be applied to reduce the number of attributes. Dimensionality reduction projects the data to fewer dimensions that retain data's fundamental attributes (Louridas & Ebert, 2016). Latent Semantic Analysis (LSA) is a common method used to reduce the dimensionality for text data. LSA analyses related concepts between the documents, the terms, and the relationships between them, then compute the similarities between the documents (Srihari et al., 2008; Kaja Zupanc & Bosnić, 2017). Then, the new matrix from LSA result is used for the clustering process. However, when LSA was implemented by using Singular Value Decomposition (SVD) in this research, the results are not improved significantly. As shown in Figure 31, the cumulative proportion of singular value does not achieve more than 70%, while the critical value to decide the optimal number of components retained should be in the range 70% until 90% (Rahim, 2017). Moreover, the number of documents for this dataset is less than 50, and the word vectors are not considered as high dimensionality data. Therefore, the original TF-IDF matrix is used without any dimensionality reduction.

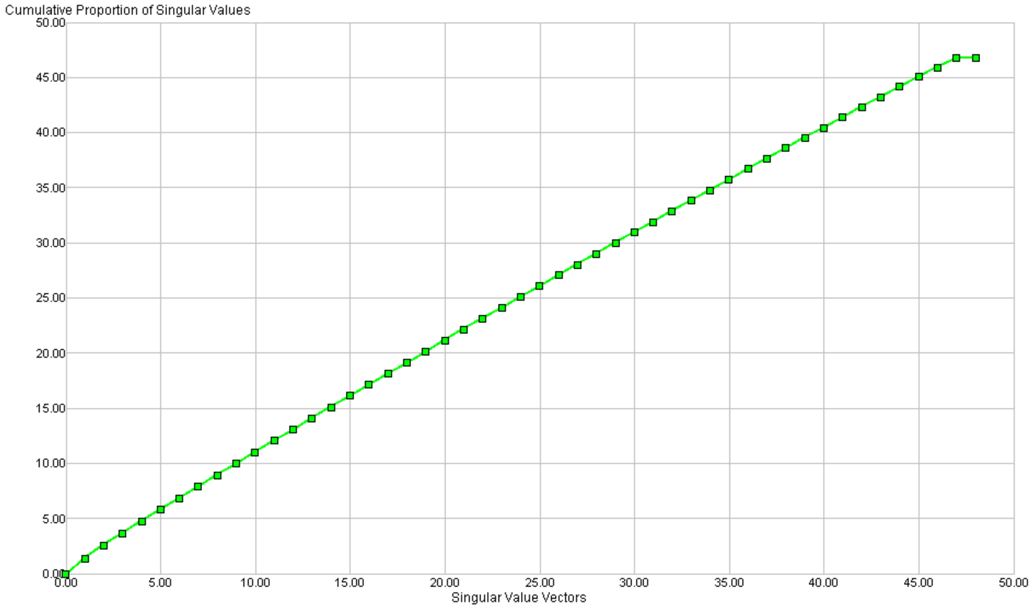


Figure 31 Scree plot of SVD Results

This study uses X-Means algorithm for the clustering. To simplify the process, Loop Parameters operator (the X-Means (2)) is used. This operator contains the X-Means, Cluster Distance, and Silhouette operator. Because X-Means operator is inside the Loop Parameters, the combination of several *k* min values and the numerical measure for the operator can be specified. The benefits of using this operator are the researcher did not have to change the parameters and run the same process repeatedly, the operator can produce all required results at once, and the optimal result can be analyzed faster. The figures of how to set the parameter setting are available in Appendix A.3.

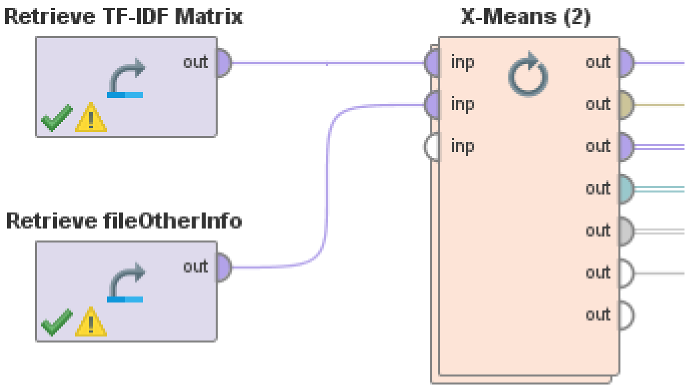


Figure 32 Clustering Implementation

The researcher tested *k* value from 2 to 10 and Euclidean, Manhattan, Correlation, and Cosine Distance as the numerical measures. Previous works in clustering used these similarity measures (Amine, Elberrichi, & Simonet, 2010; Bafna et al., 2016; Huang, 2008; Mishra, Agrawal, & KumarPatidar, 2012). However, after experimenting with those parameters and measurement index, Euclidean and Manhattan similarity shows positive result compared to the others. As a result, this research uses both similarity measures for further analysis.

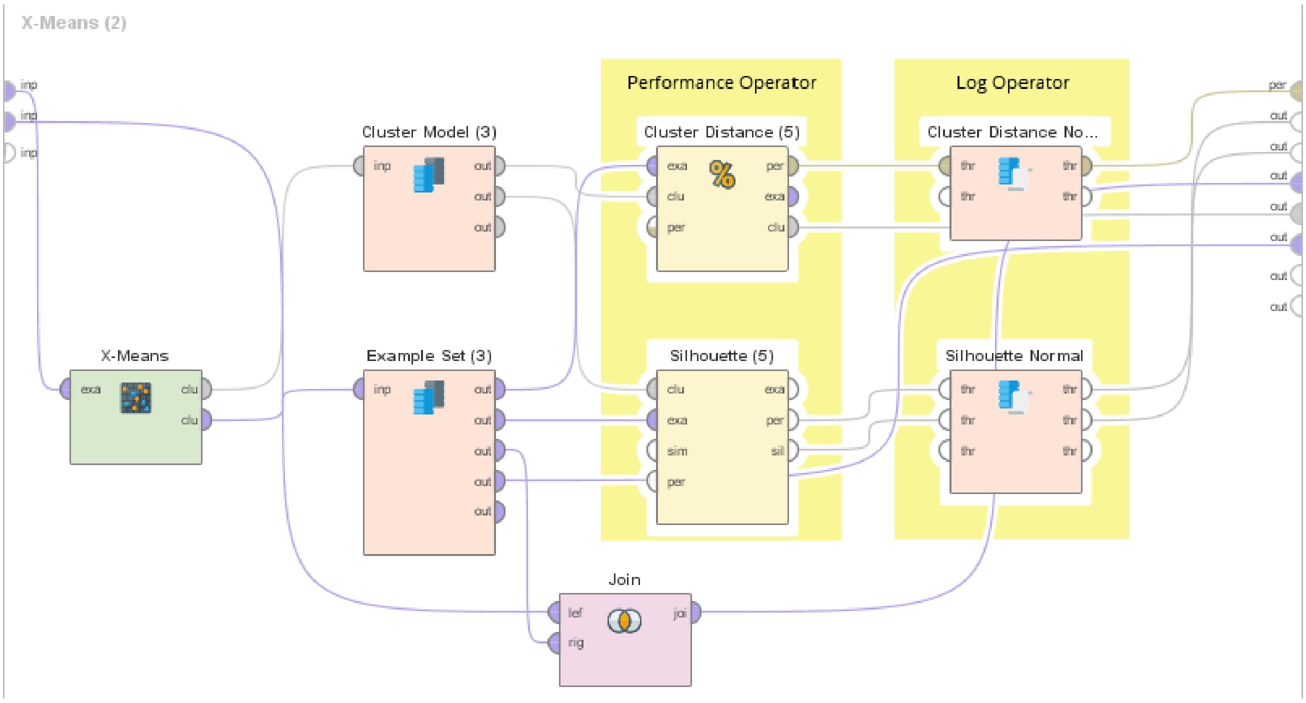


Figure 33 Process Inside X-Means (2) Operator

- Cluster Distance Performance operator generates a DB index for each cluster. DB index is useful to determine optimal *k* value.
- Silhouette Performance operator measures the quality of a cluster. It is a third-party plugin and cannot be downloaded directly from the marketplace. The link to download the java plugin and the instructions on how to install it in the RapidMiner can be found on this page: https://idealbook.wordpress.com/2015/12/08/adding-silhouette-to-rapidminer/
- Log operators (the Cluster Distance Normal and the Silhouette Normal) displays the results of Cluster Distance and Silhouette operator, which measures the performance of the clustering. The Log operators also show *k* value and numerical measures used for current iteration. Because Log results cannot be stored into an object, the results are saved manually to an Excel file for the analysis.
- Join operator joins the example sets of filename and ID with the clustering result, so it is easier to read the output because clustering does not produce the result based on the same order as in its input.

CHAPTER 5 – RESULTS

This chapter reports the results of the demonstration and evaluation phase in DSRM. The demonstration phase was conducted in an experiment described in Chapter 4. On the other hand, the evaluation was performed by inviting three lecturers and giving a presentation to them about the proposed method. At the end of the presentation, questionnaires were handed out to the lecturers to ask about their opinion of the proposed method. The concept of Unified Theory of Acceptance and Use of Technology (UTAUT) by Venkatesh et al. (2003) becomes the reference to do the evaluation phase.

5.1. Experiment Results

This section contains the results of each question type for methods implementation in RapidMiner and the recommendations or suggestions on how a question or the student answer should be formulated and formatted, or how a method should be implemented within an automated grading system to simplify the task of grading it automatically.

5.1.1. Mention a number of examples and explain

This section describes the results of two methods used to assess a question that asks the student to mention and explain a number of examples. In the experiment, there are three questions used as a comparison for this kind of question.

E-commerce 2018 exam
First Question

Please give two examples of **benefits** and two examples of **limitations** of Blockchain technology.

Business Case Development for IT-project (BCD4IT) 2018 exam
Second Question

a. What are in your opinion three **strong points** of this business case?
b. What are in your opinion three **weak/improvement points** of this business case?

E-commerce 2018 exam
Third Question

Forbes listed Trivago as the best app that saves you money. Through the listing, Forbes offers increased visibility to Momondo. What **value object** would Forbes receive in return, from Momondo? Explain your answer.

The first two questions are similar in term of the availability of how many examples/points and more than one category being asked, whereas the last question does not require the students to give a particular number of the value object. In the last question, the student can give one or more examples with a valid argument.

a. Counting Number of Examples

Based on the method implemented in this study, the table below shows how accurate it detects the number of examples for each category in all answers. Detail results of counting examples for each answer can be found in Appendix B.1. and Appendix B.2.

Table 6 Accuracy Results of Counting Number of Examples and Categories

	Benefits/Strong Points	**Limitations/Weak Points**
First Question	43.33%	63.33%
Second Question	82.98%	74.47%

This method cannot be applied for the third question because it is difficult to identify the value object without numbering or bullet. Conjunction words, such as and, furthermore, or additionally, could be used to detect a value object, but it might capture incorrect results since those words are also used to express a different idea and not another example. Moreover, in this case, most students use narrative or comma (,) to explain the examples and it is complicated to discover that information automatically.

The first question has lower accuracy than the second because of incompatible answers format with the method. This problem also occurs in the answers of the second question, but the frequency is less than the first one. Consequently, the results for the second question are higher than the first question. The rest of this section explains more about the incompatible format.

The first case is some students write the answer in a separate line for the short answer and the example. For example,

> **Benefits**
> - Secured transfer of any virtual asset with decentralized trust
> - Example: EPR (Electronic Patient Record) on a blockchain so medical data OR the whole logistics / supply chain process on a blockchain
> - Smart contracts (a new way to sign and automatically execute contracts)
> - Example: Imagine shipping a container with flowers from Vlaardingen to Argentina, the flowers have to stay below 3 degree Celsius. If an IoT sensor measures 5 degrees, a smart contract is triggered and automatically refunds the shipping fee to the sender (or if the sender also signed up for an insurance, then the whole value is refunded, all instant!)
>
> **Limitations**
> - Some of the consensus models draft enormous amounts of energy which makes it a highly unsustainable technology.
> Example: Bitcoin miners energy usage
> - Accessibility is still a major problem, in other words, integration with existing systems / applications. This is one of the main results from a research by Forrester.
> Example: Integrate Blockchain with existing supply chain management software, this will be very difficult.

By looking directly with our eyes, it is obvious that the students mention only two benefits and two limitations. However, because the method considers a new line as a new example, the methods return four benefits and four limitations for this answer.

Second, there are also students who write their answer in a paragraph instead of lists. This writing style leads to miscalculation. In the following example, the method generates only one category (the disadvantage) with one example, while there are advantage and disadvantage with two examples for each.

> Using a Blockchain in a supply chain for the Logistics may have **advantages** such as, the Blockchain helps to keep all the information from the begging of this process, so many mistakes can be avoided. Also, Blockhains are more difficult to be hacked, a vital issue when there is a huge amount of data. On the other hand, the **disadvantages** of Blockchains in a supply chain need money and lot of time to become.

Another situation is the student uses different words for the category than mentioned in the question. The current method specifies the category and its synonyms, such as benefits or advantages, and limitation, downside, or disadvantage. The drawback of this method is the teacher should define the keyword and the synonym(s) manually. Thus, the method misses the examples if the student writes another word for the category and yields incorrect calculation. For example,

> **Good**:
> (1) Auditability, due to the transactions being known and distributed. Tampering is made impossible.
> (2) Efficiency and Speed, due to the decentralized nature architecture.
>
> **Limitations**:
> (1) Wasted resources, due to the amount of computation needed when for instance mining new BitCoins. This already has a huge CO2 footprint while (arguably) only little is produced to compensate for this waste.
> (2) Anonymity, as the owner of a wallet is hard to track.

Although the student already writes in separate lines for each example, because he/she used "Good" for benefit category, the method does not identify it. Synonym checking could be implemented in the future to handle this type of case. Defining a template for the answer can also be the solution to avoid this mistake.

b. Content Checking

Typical measurements to assess the performance of an automated grading machine are about agreement (reliability) between human grader(s) and machine scoring, which can be measured in accuracy, Cohen's kappa, Quadratic Weighted Kappa (QWK), Pearson's *r* correlation coefficient, and agreement rate (exact or adjacent agreement).

Cohen's kappa and QWK are inter-rater statistic measurement to assess the degree to which each marker agreed with the answers (Butcher & Jordan, 2010). Kappa value takes into account the possibility of random agreement (Nakamura, Murphy, Christel, Stevens, & Zollman, 2016; Shermis, 2014). While kappa considers all disagreements the same, QWK treats the weighted distance between the pairs of ratings so that two ratings that far apart have a more negative impact on the measure than two ratings that are not exact agreements but are closer (Shermis, 2014). The acceptable value for kappa measurement is at least 0.70 (Ramineni & Williamson, 2013; Williamson, Xi, & Breyer, 2012).

Pearson's *r* correlation coefficient is used to measure the strength of a linear association between two variables (Jin et al., 2017). A good *r*-value should not be below 0.70 (Ramineni & Williamson, 2013; Williamson et al., 2012).

There are two types of agreement, which are exact and adjacent agreement. Exact, or perfect, the agreement is when the machine grader gives exactly the same score with a human grader, while adjacent agreement system does not mark as similar as the human grader, but the score is adjacent by one or two points (Fazal, Hussain, & Dillon, 2013). Some studies report for each exact and adjacent agreement separately, but some combine both of them. The term of agreement and accuracy are interchangeable. The greater the value is, the better the performance of the system is.

The table below presents the threshold value of the measurement (Burrows et al., 2015; Ramineni & Williamson, 2013).

Table 7 Threshold Value of Measurement

Measurement	Threshold value
Kappa	< 0.40 (poor) 0.40 – 0.75 (fair to good) ≥ 0.75 (excellent)
r	0 – 0.10 (none) 0.10 – 0.30 (small) 0.30 – 0.50 (medium) 0.50 – 1.00 (large)

In this study, since the objective of this thesis is not to give a grade to the answers, it is difficult to use similar measurement metrics as the metrics require the predicted score from the system to be able to compare the result. Therefore, this study uses a measurement metric of accuracy and correlation for the different context. This metric is used to derive the diversity of the answer. The higher value of accuracy means the answers might be similar to each other. To know how many answers are different than the others, this study uses the confusion matrix. On the other hand, the higher value of correlation could indicate that the answers are more similar to each other.

This study tests Naïve Bayes, SVM, and KNN algorithms and different parameters were explored. After several experiments, SVM achieves the best performance compared to other algorithms because it can distinguish better between strong and weak categories. Detail results for other algorithms are available in Appendix B.3. and B.4.**B.4.** The results summary for the content checking of all answers using SVM algorithm is displayed in Table 8.

During the experiment, the method does not work either for the third question because there is only one category in the question. The proposed method needs at least two categories. Therefore, the implemented method is not suitable with a question that asks one category only.

Table 8 Results of Content Checking

	Accuracy	Correlation
First Question		
Training Set	90.91%	0.818
Evaluation Set	88.89%	0.778
Second Question		
Training Set	92.42%	0.858
Evaluation Set	85.71%	0.714

Table 9 Confusion Matrix of Training Set Second Question

	true strong	**true weak**	**class precision**
pred. strong	28	0	100.00%
pred. weak	5	33	86.84%
class recall	84.85%	100.00%	

Table 10 Confusion Matrix of Testing Set Second Question

	true strong	**true weak**	**class precision**
pred. strong	12	2	85.71%
pred. weak	2	12	85.71%
class recall	85.71%	85.71%	

The overall results for the first and second question are high. The accuracy results for the evaluation test is above 85%. The correlation values are also high and indicate quite a strong relationship between each answer.

From the confusion matrix of the second question, in the training set, five answers of the strong category are predicted as weak. In this question, one weak point of the business case is the executive summary. When looking at the student answers, the model recognizes several answers, in the strong category, that mention executive summary should belong to the weak category. This signifies that the model learns quite well from the data. The detail of the results can be seen in Appendix B.5. and B.6.

The training set overall results of BCD4IT is higher than e-commerce dataset. This is most likely because there are more training data in the BCD4IT exams (33 samples) than in e-commerce exam (21 samples). However, the result of the evaluation set in e-commerce is higher than BCD4IT exam. It is presumably because the answers in e-commerce exam are more well-defined than in BCD4IT exam. If we take a look of the question, the e-commerce exam asks about benefits and limitations of Blockchain technology, which have a set of obvious and globally accepted answers. Contrarily, the question of BCD4IT asks about the opinion of the students of strong and weak points of the business case. As all students do not have the same opinion, there could be an unexpected answer in evaluation data that are not captured in the training data, and it affects the performance results because the classifier receives a relatively new answer.

Furthermore, the training data in e-commerce exam includes a text file consists a set of possible correct answers. This study uses student answers as the training data to determine the content of an answer.

Consequently, the results of the training process only capture the terms mentioned in the student answers, and other correct answers that are not written in the answers are excluded. By preparing this answer file, the classifier model retains all possible correct answers, and when uncommon correct answer appears in the evaluation data, the model can classify it better.

Because only BCD4IT exam that has scores for the answer, the ranking is performed for question 2 only. The ranking result is compared with the original score. From the training results, the sum of confidence does not sort the answer from the highest score to the lowest score. However, the answer with a lower original score (around 6 and 7) are all placed below the average confidence. This could be an indication that the combination of the SVM algorithm and confidence ranking does decent work in separating the good and bad answer. The results can be seen in detail in the Appendix B.7. and B.8.

c. Recommendations

From the experiment results, there are several recommendations for this type of question.

- Always state the number of examples being asked. People most likely write the answer in lists when the question specify how many examples it requires. Having a fixed amount is also helpful to reduce the grading workload and to ensure fair grading.
- In some cases, several people write in paragraph style instead of lists. A template could be prepared to make a more standardized format answer for all students and assure no miscalculation or missed answer. One example of the template could be simply defined like below.

Two benefits:
1. ….
2. ….

Two limitations:
1. …
2. …

- Construct a question in a format like the second question. All answers of the second question are marked as sufficient answers because each category and the examples in the answers are written using bullets, numbering, or in a new line. The question format like the following could influence the students to write their answers in a similar style as the question.

a. Mention <the number> examples of <category 1> and explain.
b. Mention <the number> examples of <category 2> and explain.

- A file contains a set of possible correct answers could be beneficial in content checking. For a question that asks about general theory or prevalent concept and has particular answers, the probability of the students giving different answers is small. This file is included as training data and could improve the model performance. However, it is important to remember that having a key answer file that can capture all possible answers is difficult, especially for a case study question.

5.1.2. Ask opinion from the students

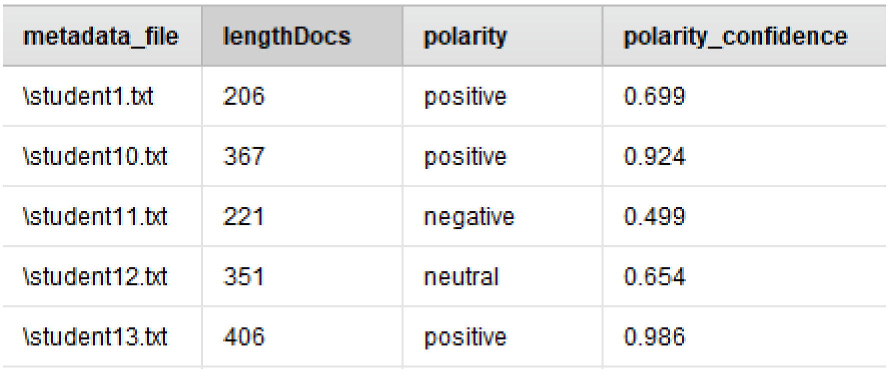
This section explains the result implementing methods to examine a question that asks opinion from the students about a statement. In this study, the question is taken from Business Case Development for IT-project 2018 exam, question 1.

> Give your view on the proposition below:
>
> *The rise of Internet of Things (IoT) and Big Data will significantly impact the field of Business Case Development.*
>
> Illustrate your position with an example.

There are two methods implemented in this type of question: sentiment analysis and clustering.

a. Sentiment analysis

Sentiment analysis from AYLIEN and counting the answers length generates a result like in Figure 34. The polarity, polarity_confidence, and lengthDocs are sent to clustering process. The full results of sentiment analysis can be seen in Appendix B.9.



metadata_file	lengthDocs	polarity	polarity_confidence
\student1.txt	206	positive	0.699
\student10.txt	367	positive	0.924
\student11.txt	221	negative	0.499
\student12.txt	351	neutral	0.654
\student13.txt	406	positive	0.986

Figure 34 Sentiment Analysis Results using AYLIEN Text Analytics

b. Clustering

This research used X-Means and Euclidean and Manhattan similarity measure in the clustering algorithm, DB index and Silhouette index as the measurement. Lower DB index and higher Silhouette value are considered a good cluster. Table 11 below list the top five DB index and Silhouette index.

Table 11 Top 5 of DB and Silhouette Index Results

Number of Clusters	Similarity Measure	DB Index	Number of Clusters	Similarity Measure	Silhouette Index
10	Manhattan Distance	0.311	7	Manhattan Distance	1.332
7	Manhattan Distance	0.334	10	Manhattan Distance	1.321
9	Manhattan Distance	0.350	9	Manhattan Distance	1.312
9	Euclidean Distance	0.350	9	Euclidean Distance	1.312
7	Euclidean Distance	0.359	6	Euclidean Distance	1.295

Manhattan Distance achieves a better result than Euclidean Distance in this study. Moreover, the cluster of 7 and 10 have an interchangeable ranking in DB index and Silhouette index. Silhouette index for each cluster is examined to determine which one is the best, and all clusters in 7 clusters obtain Silhouette index more than 0.500. Therefore, the optimal cluster for this dataset is 7 clusters with Manhattan Distance. Table 13 shows the summary of the cluster.

Table 12 Silhouette Index for 7 and 10 Clusters

Cluster	Silhouette Index	Cluster	Silhouette Index
0	0.576	0	0.577
1	1.000	1	1.000
2	0.522	2	0.691
3	0.725	3	1.000
4	0.800	4	0.781
5	1.000	5	1.000
6	0.654	6	0.497
		7	0.577
		8	0.727
		9	1.000

Table 13 Summary of Clustering Results of 7 Clusters with Manhattan Distance

Cluster	Total File	Silhouette Index	Average Polarity Confidence	Average Answer Length (words)
0	18	0.576	0.719	191
1	1	1.000	0.539	539
2	3	0.522	0.948	385
3	9	0.725	0.705	105
4	7	0.800	0.681	262
5	1	1.000	0.693	450
6	8	0.654	0.790	329

By looking at the human score, the result does not group the answers well. For example, although all files in cluster 6 originally have score 10, the index is the lowest among the others. The reason is the features used for the clustering are documents length, TF-IDF matrix, and the polarity from sentiment analysis. Therefore, the result should be analyzed based on those features.

Table 14 Clustering Results Compared to Human Score

Cluster	Answer File Number	Original Human Score
0	1,11,16,18,19,21,22,24,29,3,38,39,41,43,5,7,8,9	4,9,10,10,7,10,7,10,4,10,8,6,10,8,7,8,8,7
1	31	10
2	10,13,2	10,10,10
3	15,17,26,28,30,34,4,46,47	4,2,7,8,8,6,6,8,7
4	14,20,23,35,37,44,45	8,10,10,6,8,10,8
5	25	7
6	12,27,32,33,36,40,42,6	10,10,10,10,8,10,8,8

Regarding the polarity, cluster 0, 2, and 3 consists of all sentiments, cluster 5 consists of 2 neutral and positive, and cluster 1, 4, and 6 consists of one sentiment only. The Figure 35 below records the distribution of all sentiments in each cluster. Because the more positive results are obtained than the negative and neutral ones, all clusters tend to have a more positive answer. It appears that the cluster does not combine all sentiments in one group. Other features might have more influence than the sentiment.

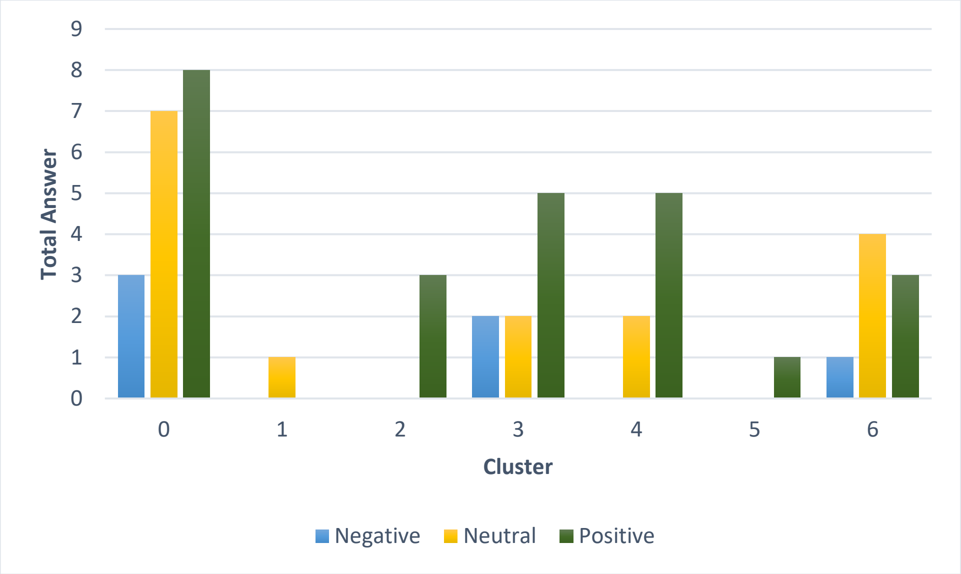


Figure 35 Polarity Distribution of Each Cluster

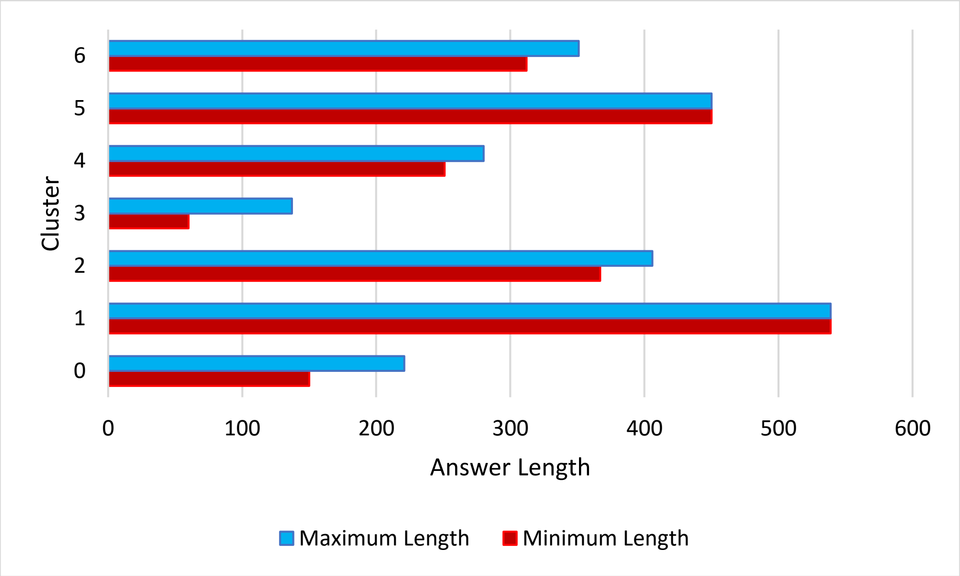


Figure 36 Maximum and Minimum Answer Length of Each Cluster

After analyzing the answer length, the answers are grouped very well based on this feature. Figure 36 shows the minimum and maximum answer length in increasing order. Cluster 3 have shorter answers, while the longest answers, with more than or equal with 450 words, are placed into the separate cluster (cluster 1 and 5). Cluster 1 and 5 are not gathered in the same cluster because it has different main points. Cluster 1

(Student31) talks about how the rise of big data and IoT change the business case development process. The student gives comparisons how those technologies give benefits to business and influence decision making by creating a data-driven business case. On the other hand, cluster 5 (Student25) only discusses how big data and IoT change business and daily life. It does not provide any relation of the technologies to the development of the business case.

When comparing the answer length and the score, longer answers tend to give clearer explanation than the shorter one. From cluster 6 until cluster 1, 3 clusters have score 8 and 10. However, in some cases, the shorter answer might elaborate better. For example, cluster 5 has the second longest answer, but several answers in cluster 0 and 4 can describe better in shorter explanation.

c. Recommendations

Grading opinion is more difficult than the previous type because every student has their own perspective and there is no right or wrong answer as long as the students have a valid argument. Therefore, there are no common properties for answers that express a personal opinion. In this study, the length of the answer determines the cluster group. This can be a possible solution to group the answer from the shorter to longer answer, so the human grader can choose from which group he/she can grade the answer. However, it should be remembered that the answer length does not determine the quality of the answer.

Regarding the tools used, Sentiment Analysis from AYLIEN might not produce the best result. For example, this answer obtains negative polarity, when the student actually agrees with the statement.

> For another course, New Production Concepts, I have been writing an article about the (ongoing) development of Industry 4.0. The rise of Internet of Things (IoT) and Big Data are both important for this (ongoing) development, because the Industrial Internet of Thins and Big Data and analytics are both one of the nine pillars that are transforming industrial production (Source: BCG). The rise of these technologies will have an impact on operations management and therefore it will also have an impact on Business Case Development. This is because of the definition of Operations Management: Operations management (OM) can be defined as "the activity of managing resources that create and deliver services and products" {Source: Slack et al.}. Business Cases can be used as input resources to create and deliver services and products. Therefore, the rise of (industrial) Internet of Things and Big Data will have a significant impact on the field of Business Case Development.
>
> Another example of the impact of the rise of big Data is that Big Data can be used in the future to better predict the options in Business Cases, which will probably have a positive impact on Business Case Development.

Perhaps in the future, a personalized sentiment analysis algorithm could be developed by specifying the customized keywords as the indications for each of the sentiment. Implementing a new algorithm and keywords could predict more suitable sentiment and thus improve the performance of the method.

Another alternative besides using clustering and sentiment analysis is by giving options for students to state directly in their answer in which position they are, then group the answers based on the options and assessed the answers by their groups. For example, in this question, the instruction could be modified as

“Illustrate your position (agree/disagree/neither agree nor disagree) with an example.”

The student can answer by writing their preference in the first sentence and the argument in the next sentences.

"I <preferred viewpoint> with the statement. <the reasoning behind the selection>"

5.2. Evaluation Result

This section reports the result of the evaluation phase in DSRM. The evaluation was performed by inviting three lecturers in the University of Twente to the presentation about the method and gathering their opinion using a questionnaire. The questionnaire was created using Unified Theory of Acceptance and Use of Technology (UTAUT) by Venkatesh et al. (2003).

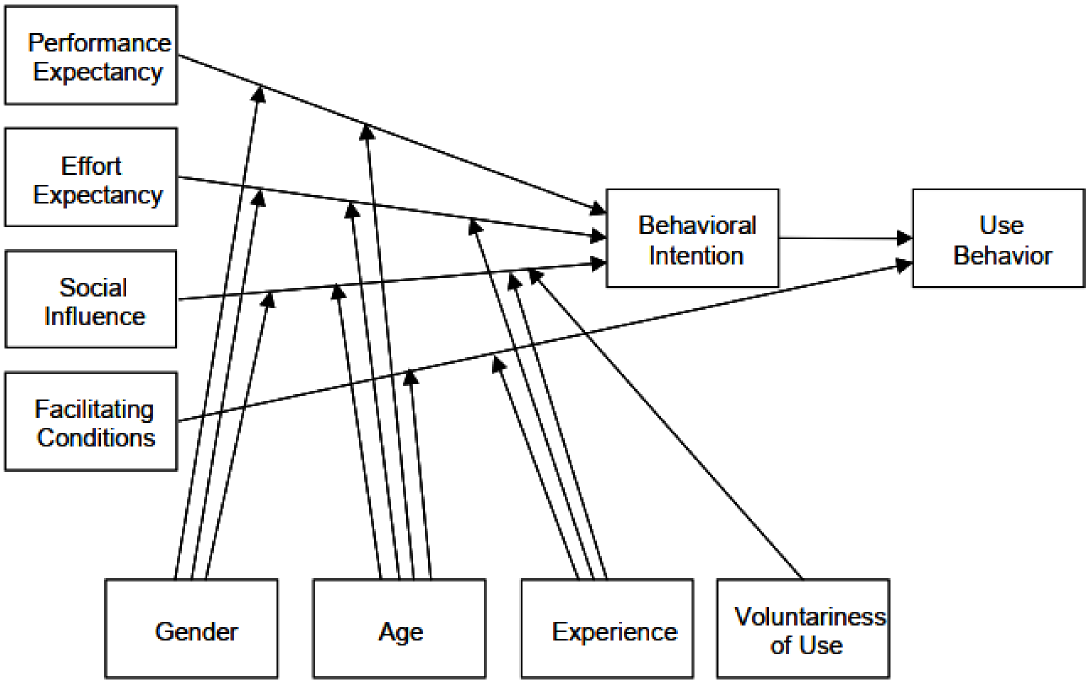


Figure 37 UTAUT Research Model (Venkatesh et al. 2013)

This study selects the UTAUT model for the evaluation process because it is a comprehensive and unified model from eight user acceptance models. The model consists of four direct determinant constructs of user acceptance and behavior: performance expectancy, effort expectancy, social influence, and facilitating conditions; while the indirect determinant factors are attitude toward using technology, self-efficacy, and anxiety. There are key moderators in this model, name gender, age, experience, and voluntariness of use. However, this study does not include the key moderators as they are not determinant variables of user acceptance and behavior. Table 15 explains all constructs in the UTAUT model.

Table 15 Constructs Summary in Estimating UTAUT

Construct Name	Definition	Root Constructs	Items
Direct Determinant			
Performance expectancy	The degree to which an individual believes that using	Perceived usefulness (TAM/TAM2 and C-TAM-TPB), extrinsic motivation	U6: I would find the system useful in my job. RA1: Using the system enables me to accomplish tasks more quickly.

	the system will help him or her to attain gains in job performance	(MM), job-fit (MPCU), relative advantage (Gupta et al.), and outcome expectations (SCT)	RA5: Using the system increases my productivity. OE7: If I use the system, I will increase my chances of getting a raise.
Effort Expectancy	The degree of ease associated with the use of the system	Perceived ease of use (TAM/TAM2), complexity (MPCU), and ease of use (Gupta et al.)	EOU3: My interaction with the system would be clear and understandable. EOU5: It would be easy for me to become skillful at using the system. EOU6: I would find the system easy to use. EU4: Learning to operate the system is easy for me.
Social Influence	The degree to which an individual perceives that important others believe he or she should use the new system	Subjective norm (TRA, TAM2, TPB/DTPB, and C-TAM-TPB), social factors (MPCU), and image (Gupta et al.)	SN1: People who influence my behavior think that I should use the system. SN2: People who are important to me think that I should use the system. SF2: The senior management of this business has been helpful in the use of the system. SF4: In general, the organization has supported the use of the system.
Facilitating Conditions	The degree to which an individual believes that an organizational and technical infrastructure exists to support the use of the system	Perceived behavioral control (TPB/DTPB, C-TAM-TPB), facilitating conditions (MPCU), and compatibility (Gupta et al.)	PBC2: I have the resources necessary to use the system. PBC3: I have the knowledge necessary to use the system. PBC5: The system is not compatible with other systems I use. FC3: A specific person (or group) is available for assistance with system difficulties.
Indirect Determinant			
Self-efficacy	The judgment of one's ability to use the system to accomplish a particular job or task	Self-efficacy (SCT)	I could complete a job or task using the system… SE1: If there was no one around to tell me what to do as I go. SE4: If I could call someone for help if I got stuck. SE6: If I had a lot of time to complete the job for which the software was provided. SE7: If I had just the built-in help facility for assistance.
Anxiety	Evoking anxious or emotional reactions when it comes to performing a behavior	Anxiety (SCT)	ANX1: I feel apprehensive about using the system. ANX2: It scares me to think that I could lose a lot of information using the system by hitting the wrong key. ANX3: I hesitate to use the system for fear of making mistakes I cannot correct. ANX4: The system is somewhat intimidating to me.

Attitude toward using technology	An individual's overall affective reaction to using a system	Attitude toward behavior (TRA, TPB/DTPB, C-TAM-TPB), intrinsic motivation (MM), affect toward use (MPCU), and affect (SCT)	A1: Using the system is a bad/good idea. AF1: The system makes work more interesting. AF2: Working with the system is fun. Affect1: I like working with the system.
Behavioral Intention to Use	A person's perceived likelihood or subjective probability that he or she will engage in a given behavior ((CHIRr))		BI1: I intend to use the system in the next <n> months. BI2: I predict I would use the system in the next <n> months. BI3: I plan to use the system in the next <n> months.

Based on the definition, this study uses performance expectancy (PE), effort expectancy (EE), facilitating conditions (FC), self-efficacy (SE), attitude toward using technology (ATT), and behavioral intention to use (BIU) as the aspects to measure in the questionnaire. The questionnaire design is available in Appendix C.

The following part discusses the result from the questionnaire. Table 16 presents the descriptive statistics (the summary of the sample and the measures) of the survey. The explanation of each data is as follow:

- N is the number of participants
- N1, N2, and N3 denote the response of participant 1, 2, and 3 for the related statement
- Min is the minimum score given by the participant for the related statement in the questionnaire
- Max is the maximum score given by the participant for the related statement in the questionnaire
- Sum is the total score given by the participant for the related statement in the questionnaire
- Mean is the average score given by the participant of the related statement in the questionnaire
- STDEV is the standard deviation of the scores given by the participant. Standard deviation measures the amount of dispersion of a set of data values. Higher standard deviation indicates the data are distributed into a wider range, and the lower value implies the data leans toward the mean.

Table 16 Descriptive Statistics of The Survey Results

Statement	N	N1	N2	N3	Min	Max	Sum	Mean	STDEV
PE-1	3	4	4	3	3	4	11	3.67	0.58
PE-2	3	4	4	2	2	4	10	3.33	1.15
PE-3	3	5	5	3	3	5	13	4.33	1.15
EE-1	3	3	3	4	3	4	10	3.33	0.58
EE-2	3	3	4	4	3	4	11	3.67	0.58
EE-3	3	4	4	4	4	4	12	4	0
FC-1	3	5	3	3	3	5	11	3.67	1.15
FC-2	3	4	5	4	4	5	13	4.33	0.58
FC-3	3	4	4	4	4	4	12	4	0
SE-1	3	3	2	2	2	3	7	2.33	0.58
SE-2	3	4	3	2	2	4	9	3	1
ATT-1	3	4	5	2	2	5	11	3.67	1.53
ATT-2	3	3	5	4	3	5	12	4	1

ATT-3	3	4	2	4	2	4	10	3.33	1.15
BIU-1	3	5	4	2	2	5	11	3.67	1.53
BIU-2	3	4	4	2	2	4	10	3.33	1.15
BIU-3	3	5	2	2	2	5	9	3	1.73
Average PE								3.78	0.96
Average EE								3.67	0.38
Average FC								4	0.58
Average SE								2.67	0.79
Average ATT								3.67	1.23
Average BIU								3.33	1.47

The participants answer the questionnaire by giving their preference for each statement from 1 to 5, where the tendency from the lower score until the higher score ranges from strongly disagree to strongly agree. Figure 38 illustrates the summary of mean and standard deviation result. From this chart, 14 out of 17 statements have mean value above 3. This indicates the participants tend to have a positive response about the proposed method.

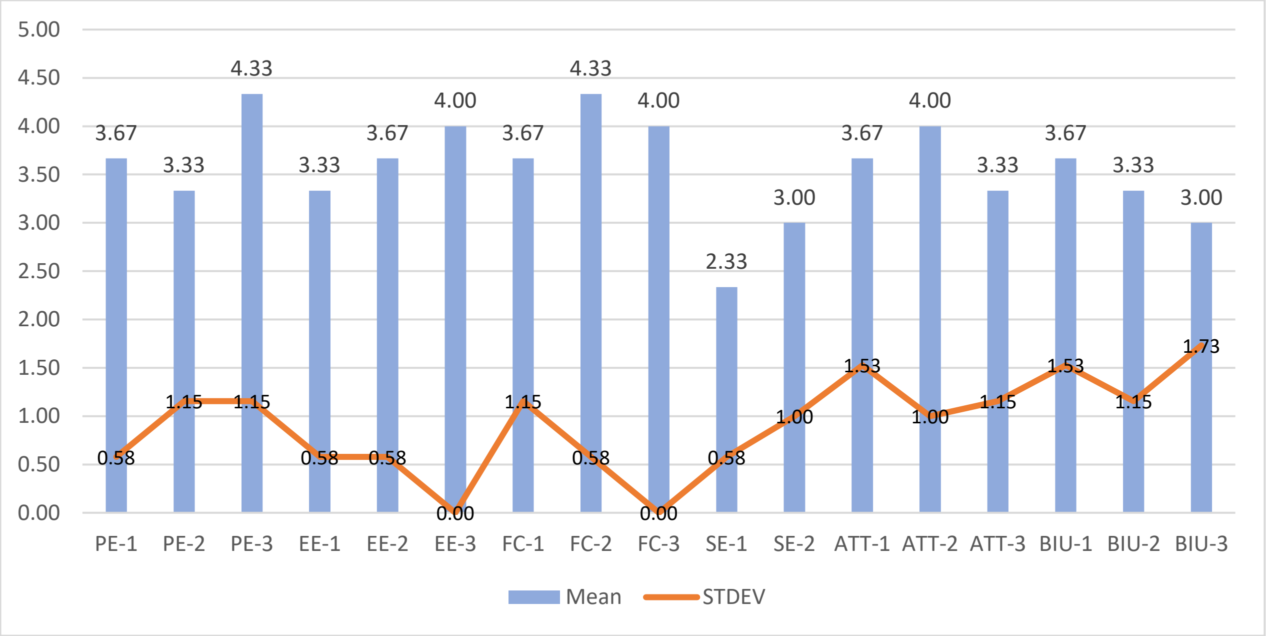


Figure 38 Descriptive Statistics of The Survey

The lowest mean is 2.33 for statement 1 in self-efficacy measurement, and the highest score is 4.33 for statement 3 in performance expectancy and statement 2 in facilitating condition. The most dispersed responses are statement 3 in behavioral intention to use, while all participants give the same score in the third statement of effort expectancy and facilitating conditions. The detail of each construct is elaborated in the following parts.

5.2.1. Performance Expectancy

Performance expectancy is the degree to which an individual believes that using the system will help him or her to attain gains in job performance. The average of this construct is 3.78, and the mean of each statement is above 3. Table 17 lists the statements for performance expectancy and the response from the participants.

Table 17 Performance Expectancy Survey Results

Statement		N1	N2	N3
PE-1	I would find the proposed method useful in grading an answer.	4	4	3
PE-2	Using the proposed method enables me to grade more quickly.	4	4	2
PE-3	Using the proposed method increases my productivity.	5	5	3

Participant 1 and 2 have the same responses for all statements, but participant 3 have a different opinion. The difference sentiment from participant 3 is because he/she expects the method to give a score prediction to the answer. The participant thinks that the groups of the answer could be beneficial for the grading process, but it is not significant. On the contrary, the other participants estimate there will be a lot of time-saving to grade the answers when the answer is already classified. The grading task will also be easier when similar answers are grouped. The conclusion is that most participants believe that using the proposed method will help them to improve their job performance.

5.2.2. Effort Expectancy

Effort expectancy is the degree of ease associated with the use of the system. Each statement in this construct achieves the average above 3, and the average score of this construct is 3.67. Table 18 shows the statements for effort expectancy and the response from the participants.

Table 18 Effort Expectancy Survey Results

Statement		N1	N2	N3
EE-1	My interaction with the proposed method would be clear and understandable.	3	3	4
EE-2	I would find the proposed method easy to use.	3	4	4
EE-3	It would be easy for me to become skillful at understanding the proposed method.	4	4	4

As we can see, this construct tends to have neutral and agree responses from the participants. Most participants respond with neutral for statement EE-1 probably because there is no real demonstration during the evaluation session in how to use the proposed method, hence they cannot imagine the usage of the method in real life. Statement EE-2 receives a more positive response than EE-1 since most participants thought the proposed method is simple. Additionally, statement EE-3 obtains the most positive result for this construct because all participants agree that because the proposed method is not complicated, the learning process to understand the proposed method will not take too much time. Overall, the participants agree that using the proposed method is easy.

5.2.3. Facilitating Conditions

Facilitating conditions is the degree to which an individual believes that an organizational and technical infrastructure exists to support the use of the system. The average of this construct is 4, and the mean of each statement is above 3.5. Table 19 displays the statements for facilitating conditions and the response from the participants.

Table 19 Facilitating Conditions Survey Results

Statement		N1	N2	N3
FC-1	I have the resources necessary to use the proposed method.	5	3	3
FC-2	I have the knowledge necessary to use the proposed method.	4	5	4
FC-3	The proposed method is not compatible with other systems I use.	4	4	4

Only one participant strongly agrees with statement FC-1, and the others neither agree nor disagree. These responses are assumed to be related with statement FC-3 in which all participants agree that the proposed method is not compatible with other systems they use. The reason behind the result is since the proposed method is implemented using RapidMiner, all participants think this software is mandatory to use the method. Another reason could be because in the method implementation, this study uses the answer in digital format while several courses still use paper examination. In other words, since the proposed method is not compatible with other systems that they use, all participants tend to feel they don't have the essential resources to use the proposed method.

However, all participants perceive they have the fundamental knowledge to use the proposed method because they understand the concept of text mining and machine learning. For other lecturers who don't have any knowledge about these concepts might have a different response. In conclusion, all participants believe that they have adequate prerequisite to use the proposed method.

5.2.4. Self-efficacy

Self-efficacy is a judgment of one's ability to use the system to accomplish a particular job or task. The average of this construct is 2.67 – which is the lowest among the other constructs — and the maximum mean of each statement is 3. Table 20 shows the statements for facilitating conditions and the response from the participants.

Table 20 Self-efficacy Survey Results

Statement		N1	N2	N3
SE-1	I could grade using the proposed method if there was no one around to tell me what to do as I go.	3	2	2
SE-2	I could grade using the proposed method if I could call someone for help if I got stuck.	4	3	2

The responses from participants for this construct tends to have negative to neutral feedback. The majority of the participant disagree with statement SE-1. This might be because there is no demonstration session during the presentation, so the participants do not have a clue about how to use the proposed method in a real case. Furthermore, most participants think that they could grade using the proposed method if they have someone to ask for help when they got stuck. This could be an indication that most participants are willing to use the proposed method when there is some guidance or tutorial in how to use the proposed method. In summary, the participants considered they cannot use the proposed method by themselves alone.

5.2.5. Attitude toward using technology

Attitude toward using technology is an individual's overall affective reaction to using a system. The average of this construct is 3.67, and the mean of each statement is above 3. Table 21 lists the statements for facilitating conditions and the response from the participants.

Table 21 Attitude Toward Using Technology Survey Results

Statement		N1	N2	N3
ATT-1	Grading answer with the proposed method is a good idea.	4	5	2
ATT-2	The proposed method makes grading more interesting.	3	5	4
ATT-3	Grading with the proposed method would be fun.	4	2	4

This construct has various feedback from each participant. Two participants agree that grading answer with the proposed method is a good idea because there is a benefit in grading answer with the proposed method, while another participant does not see any benefit on it. The second statement obtains the most positive result in this construct because presumably, the idea of working with automated grading is enticing, but the first participant thinks that whether using automated grading or not, the grading task will never be interesting. However, the first participant expects that the proposed method would make grading fun than in manual grading, and the second participant imagines the otherwise. The reversed feedback might be because the definition of interesting and fun for both participants is switched. Overall, all participants express a positive reaction to using the proposed method.

5.2.6. Behavioral intention to use

Behavioral intention to use is a person's perceived likelihood or subjective probability that he or she will engage in a given behavior. The average of this construct is 3.33, and the mean of each statement is equal or more than 3. Table 22 displays the statements for facilitating conditions and the response from the participants.

Table 22 Behavioral Intention to Use Survey Results

Statement		N1	N2	N3
BIU-1	I intend to use the proposed method in the future to help me grading student's answer.	5	4	2
BIU-2	I predict I would use the proposed method in the future to help me grading student's answer.	4	4	2
BIU-3	I plan to use the proposed method in the future to help me grading student's answer.	5	2	2

From the table, participant 1 implies the most positive response toward his or her intention to use the proposed method in grading student’s answer. On the contrary, the last participant signifies the least likelihood to grade using the proposed method. The assumption behind this situation is strongly determined with performance expectancy as Venkatesh et al. also stated in their experiment (Venkatesh et al., 2003). From the performance expectancy, the third participant already shows negative sentiment for each statement, while the two others present more positive feedback. Another thing to note is the inconsistency of the second participant in statement BIU-3. Although the participant agrees with statement BIU-1 and BIU-2, he or she somehow disagree with BIU-3. This result is probably because the participant has the intention to use the proposed method, but he/she will not use it soon. In the end, 2 out of 3 participants have positive result to use the proposed method.

From the evaluation results, performance expectancy becomes the strong determinant of the behavioral intention to use the proposed method. Although facilitating conditions achieves the most positive result among other constructs, the result is nonsignificant in predicting the intention (Venkatesh et al., 2003). It does not affect the participant intention to use the proposed method because it is an indirect determinant factor as previously stated. The most negative feedback is self-efficacy construct as there is a possibility that all participants think introduction session is important before using the system. It is assumed that after learning from the introduction session, the participants could have a better understanding of the proposed method because the participants have a positive response for the effort expectancy.

CHAPTER 6 – CONCLUSIONS

This chapter concludes the result of the research in main finding and discussion. This chapter also provides the limitations and the validity threats during the study. Moreover, this chapter offers several recommendations to overcome the limitations and several directions for improvements in future research.

6.1. Main Finding

This section aims to answer the main research question and all sub-questions.

SQ1: How should open questions be formulated to be useful for automated grading using text mining?

Each open question has its own characteristics. These characteristics should be remembered when creating a question so that the students can answer in the desired manner. In this study, the first question type has obvious properties, which are the categories and the amount of example requested. Categories are used to classify and check the content of the answer. Then, the availability of the second property can be used to group the answers so that the grader can assess the answer based on this result. Opinion question does not have any particular characteristics. One possible solution that might be beneficial is by asking the student to choose which position they prefer about the given idea.

SQ2: What kind of text mining and machine learning techniques are available to grade open questions?

There are various methods used to build an automated grading system. Regarding Burrows et al. (2015) study, there are five themes in automated grading system: concept mapping, information extraction, corpus-based, machine learning, and evaluation. The most common themes are machine learning and information extraction (IE) – both methods usually apply NLP techniques to prepare the data – while the least common is the evaluation. Machine learning method prepares the dataset to extract and select the features using NLP, including text mining approach, such as *n*-gram technique, stop word removal, stemming, and find synonyms, to build the scoring model using any machine learning algorithm. In IE technique, the scoring model is constructed based on matching the student answer with a pattern found from the dataset.

SQ3: How to design an algorithm based on text mining and machine learning techniques to support the automated grading of open questions?

The algorithm in this study is created based on the characteristics of the question and the answer. In the first question, text mining is used to split the answer into categories, count the number of examples, remove unnecessary words, and transform the text into a more structured form in term of document matrix. Then, the algorithm combines it with a machine learning approach to check the content of the answer and rank the answers based on the confidence value. The results are groups of answer(s) based on the examples amount and ordered answers based on the rank. For the second question, calculate the length of the answer. Then, execute sentiment analysis, generate the TF-IDF matrix, and do the clustering. The results are clusters of similar answers.

SQ4: In what way can the system performance be measured?

Selecting suitable evaluation metrics is an important task to examine the performance of the method. For type 1, the results of the proposed method are acceptable for answers with two categories and a fixed number of examples being asked. In counting the number of examples, the comparison between the real number of examples in the answers and the counting results is measured. It appears that a well-formatted answer improves the accuracy of the splitting and counting examples. For the classification technique, the performance can be measured by accuracy, confusion matrix, and correlation coefficient. Moreover, the ranking based on confidence value signifies a positive result in separating the answers.

Meanwhile, DB index and Silhouette index measures the quality of clustering result. The optimal result for the current dataset is 7 clusters using X-Means and Manhattan Distance. From the evaluation result, the participants consider the result of the clustering could be used to map to grade points and give a recommendation to the teacher about possible score they can give to the answer.

RQ: How can text mining and machine learning techniques be used for the automated grading of open questions?

There are several types of open questions in higher education level other than short answers and essays. For example, the answer in the form of table, picture or graph, and mention benefits of enterprise architecture for business and explain it in a real application. Each question type requires specific methods, hence a method to assess question A cannot be used directly to examine question B. The text mining and machine learning techniques can be utilized to find out the characteristics of an open question. The result of this process can be applied to build a better method in automated open grading.

Furthermore, the result can help lecturer to construct a test that is easily graded by the automated system. One purpose of the automated grading system is to reduce the workload in grading and produce a reliable result. By creating an exam that is suitable with the approaches implemented inside the system, the grading task becomes more effective, and the result of the grading is also more reliable.

Automated grading of open questions can employ various techniques in text mining and machine learning. Text mining procedures can prepare the answer text into structured form in matrix and machine learning can be implemented to classify or cluster the answer. An appropriate combination of text mining and machine learning methods can create a powerful and advantageous method for the automated grading system.

6.2. Discussion

Implementing an automated grading system in the educational system have the benefits and the limitations. This section discusses both of the benefits and limitations regarding the results of this study.

A pilot experiment by Wolska et al. (2014) indicates that grading grouped answers required less time than grading the answers when they are not grouped. Furthermore, the grading results of clustered answers are also as efficient as grading for individual answers. This study has a similar approach with Wolska et al. by collecting similar answers in one group. There are still no results to evaluate the grading time and the effort in this study, but the study by Wolska et al. and the results of evaluation survey in performance expectancy construct signify that proposed method might improve the performance of the lecturers in grading.

The results of each technique in this study depends on the dataset quality. The process of selecting and extracting the features for the machine learning determines the grouping and classification results. Therefore, answer pre-processing is an important step to obtain good quality data and the more reliable results.

Another downside of automated grading is the susceptibility to “gaming” behavior where the students can learn how the automated grading works and attempt to improve their score without writing the real answer (Grimes & Warschauer, 2010; Higgins & Heilman, 2014). The studies by Grimes and Warschauer (2010) and Bejar et al. (2014) shows that increasing the word count and keywords frequency can refine the score. Even writing the same correct answer can boost the score (Higgins & Heilman, 2014).

Because of these drawbacks, it might be preferable if the educational system does not depend completely on automated grading. Automated grading should be used to help the lecturers in grading and not replace the role of lecturers as the sole grader. The combination of the automated grading and human grader could complement each other. The strengths of automated grading, such as the consistent analysis and efficiency to produce a result, can reduce the time and effort of the lecturers to do the grading. On the other hand, the manual assessment from the lecturers yields to more fair and reliable results as the lecturers can evaluate the answer in a way the system cannot do, for example, by examining the correctness and the coherency of the answer.

6.3. Limitations

During the study, several limitations occurred.

1. Data limitation. The number of each dataset used in this study is relatively small than similar studies in this area. Furthermore, only BCD4IT dataset that contains a human grade. Therefore, it is difficult to compare the result of the method implementation with the real case.
2. Knowledge limitation. Due to limited time, the researcher did not have much time to learn and explore more about using the tools. RapidMiner has many features that most likely be advantageous in this experiment, but were not used fully because the author did not know about its features completely.
3. Design limitation. Other techniques can be used for analyzing text data that might improve the performance of the system. Other parameters that could affect the result also have not been examined because of the time limitation in this research. Moreover, different approaches could give a better result than the current approach.

In carrying out this research, we accounted for the following validity threats.

1. This study uses the limited dataset in two exams only. The result of the study cannot be generalized to other courses and question type.
2. The evaluation process has a small sample of teachers. The minimum sample required to have a valid result is 30 people, and the sample should be selected randomly. Therefore, the results do not represent the opinion of the whole population.
3. The current study only searches literature from four digital libraries. Some relevant studies might be missed from the search results. Furthermore, this research uses keywords that might not capture all relevant studies.

6.4. Future Work & Recommendation

Based on the limitations mentioned above, several things could be done in future works related to this study.

1. Finding more data to work with or doing experiment with other types of question. With more data, the performance of the system could be enhanced. Future research can use the same type of questions, but with more data, or trying to look at different types. There are many other types of open question, such

as a question which asks the student to fill in a table, draw an image or graph or a question which is related to the previous question.

2. Acquiring adequate knowledge to use the tools. Knowing and using potential functionalities of a tool could be helpful for research. Therefore, it will be better if the researcher(s) is already familiar with the tools they will use. Combining different type of tools is also possible if a single tool is not enough to do the study.
3. Investigating other approaches and variables that could enhance the result. For example, in this study, there is no spell-checking, while in reality, students could write incorrect spelling or grammar. Future research could implement error checking in the system design.
4. Using more databases to gather more relevant studies. Additionally, constructing the keywords by using common keywords in reliable and trusted papers or reading the content of the papers to obtain other related terms.
5. Having more sample of teachers for the evaluation process, at least 30 people. More sample yields more reliable results.

Furthermore, regarding the evaluation result, there are several ideas to improve the behavioral intention to use the proposed method in the future.

1. Generating a score mapping from the clustering and grouping results. Several people expect the function of an automated grading is to produce a score to the answer at the end. Therefore, future research could create a score mapping from the results, for example having a particular word in the answer will increase or reduce the score, or the answers in sufficient group will get a minimum score at 5.5. The score can be used by the teachers as the basis of their manual assessment to the answer.
2. Having an experiment to evaluate the performance of the teachers to grade the answers using the proposed method. The time spent and the effort needed by the teachers to perform the grading could be the measurement metrics for the experiment.
3. Creating a user interface for the proposed method. Then, giving a demonstration to the lecturers in a tutorial. Furthermore, developing a guideline for the teachers about how to use the system.
4. The proposed method processes the answers that are typed in a digital file. For future work, an implementation of image processing to capture the handwritten answers can be combined with the proposed method.

REFERENCES

(CHIRr), C. H. I. R. r. Behavioral Intention. Retrieved from https://chirr.nlm.nih.gov/behavioral-intention.php

Aggarwal, C. C., & Zhai, C. (2013). An Introduction to Text Mining. *Mining Text Data*, 1-10. doi:10.1007/978-1-4614-3223-4_1

Agrawal, R., & Batra, M. (2013). A Detailed Study on Text Mining Techniques. *International Journal of Soft Computing and Engineering (IJSCE), 2*(6), 118-121.

Ahmad, N. D., Adnan, W. A. W., Abdul Aziz, M., & Yusaimir Yusof, M. (2011). *Automating preparation of exam questions- Exam Question Classification System (EQCS)*. Paper presented at the 2011 International Conference on Research and Innovation in Information Systems.

Alickovic, E., & Babic, Z. (2015). *The effect of denoising on classification of ECG signals.* Paper presented at the 2015 25th International Conference on Information, Communication and Automation Technologies, ICAT 2015 - Proceedings.

Amine, A., Elberrichi, Z., & Simonet, M. (2010). Evaluation of text clustering methods using WordNet. *International Arab Journal of Information Technology, 7*(4), 349-357.

Bafna, P., Pramod, D., & Vaidya, A. (2016). Document Clustering: TF-IDF Approach. *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 61-66. doi:10.1109/ICEEOT.2016.7754750

Bandyopadhyay, S., & Maulik, U. (2001). Nonparametric genetic clustering: Comparison of validity indices. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, 31*(1), 120-125. doi:10.1109/5326.923275

Bin, L., Jun, L., Jian-Min, Y., & Qiao-Ming, Z. (2008). *Automated essay scoring using the KNN algorithm.* Paper presented at the Computer Science and Software Engineering, 2008 International Conference.

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education, 25*(1), 60-117. doi:10.1007/s40593-014-0026-8

Butcher, P. G., & Jordan, S. E. (2010). A comparison of human and computer marking of short free-text student responses. *Comput. Educ., 55*(2), 489-499. doi:10.1016/j.compedu.2010.02.012

Clay, B. (2001). Is This A Trick Question? A Short Guide to Writing Effective Test Questions. Retrieved from

Dang, S., & Ahmad, P. (2015). A Review of Text Mining Techniques Associated with Various Application Areas. *International Journal of Science and Research (IJSR), 4*(2), 2461-2466.

Driscoll, G., Hatfield, L. A., Johnson, A. A., Kahn, H. D., Kessler, T. E., Kuntz, D. L., . . . Williams, P. G. (1999). On-line essay evaluation system. In: Google Patents.

Errecalde, M. L., Cagnina, L. C., & Rosso, P. (2015). Silhouette + attraction: A simple and effective method for text clustering. *Natural Language Engineering, 22*(5), 687-726. doi:10.1017/S1351324915000273

Fazal, A., Hussain, F. K., & Dillon, T. S. (2013). An innovative approach for automatically grading spelling in essays using rubric-based scoring. *Journal of Computer and System Sciences, 79*(7), 1040-1056. doi:https://doi.org/10.1016/j.jcss.2013.01.021

Funk, S. C., & Dickson, K. L. (2011). Multiple-Choice and Short-Answer Exam Performance in a College Classroom. *Teaching of Psychology, 38*(4), 273-277. doi:10.1177/0098628311421329

Gnimpieba, E. Z., VanDiermen, M. S., Gustafson, S. M., Conn, B., & Lushbough, C. M. (2017). Bio-TDS: Bioscience query tool discovery system. *Nucleic Acids Research, 45*(D1), D1117-D1122. doi:10.1093/nar/gkw940

Gonzalez-Barbone, V., & Llamas-Nistal, M. (2008, 22-25 Oct. 2008). *eAssessment of open questions: An educator's perspective.* Paper presented at the 2008 38th Annual Frontiers in Education Conference.

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment, 8*(6).

Gupta, S., Ross, K. E., Tudor, C. O., Wu, C. H., Schmidt, C. J., & Vijay-Shanker, K. (2016). miRiaD: A Text Mining Tool for Detecting Associations of microRNAs with Diseases. *Journal of Biomedical Semantics, 7*(1). doi:10.1186/s13326-015-0044-y

Gutierrez, F., Dou, D., Martini, A., Fickas, S., & Zong, H. (2013). *Hybrid ontology-based information extraction for automated text grading.* Paper presented at the Proceedings - 2013 12th International Conference on Machine Learning and Applications, ICMLA 2013.

Hasanah, U., Permanasari, A. E., Kusumawardani, S. S., & Pribadi, F. S. (2016). *A review of an information extraction technique approach for automatic short answer grading*.

Higgins, D., & Heilman, M. (2014). Managing what we can measure: Quantifying the susceptibility of automated scoring systems to gaming behavior. *Educational Measurement: Issues and Practice, 33*(3), 36-46. doi:10.1111/emip.12036

Hladka, B., & Holub, M. (2015). A Gentle Introduction to Machine Learning for Natural Language Processing: How to Start in 16 Practical Steps. *Language and Linguistics Compass, 9*(2), 55-76. doi:10.1111/lnc3.12123

Huang, A. (2008). *Similarity Measures for Text Document Clustering.* Paper presented at the New Zealand Computer Science Research Student Conference 2008, Christchurch, New Zealand.

Husain, H., Baisb, B., Hussain, A., & Samad, S. A. (2012). How to Construct Open Ended Questions. *Procedia - Social and Behavioral Sciences, 60*, 456-462. doi:10.1016/j.sbspro.2012.09.406

Ivanović, M., & Radovanović, M. (2015). *Modern machine learning techniques and their applications.* Paper presented at the Electronics, Communications and Networks IV - Proceedings of the 4th International Conference on Electronics, Communications and Networks, CECNet2014.

Jayashankar, S., & Sridaran, R. (2017). Superlative model using word cloud for short answers evaluation in eLearning. *Education and Information Technologies, 22*(5), 2383-2402. doi:10.1007/s10639-016-9547-0

Jin, C., & He, B. (2015). *Utilizing latent semantic word representations for automated essay scoring*.

Jin, C., He, B., & Xu, J. (2017) A study of distributed semantic representations for automated essay scoring. In*: Vol. 10412 LNAI* (pp. 16-28).

Joiner, J., Yoshida, Y., Vasilkov, A. P., Schaefer, K., Jung, M., Guanter, L., . . . Belelli Marchesini, L. (2014). The seasonal cycle of satellite chlorophyll fluorescence observations and its relationship to vegetation phenology and ecosystem atmosphere carbon exchange. *Remote Sensing of Environment, 152*, 375-391. doi:10.1016/j.rse.2014.06.022

Jović, A., Brkić, K., & Bogunović, N. (2014). An overview of free software tools for general data mining. *37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1112-1117. doi:10.1109/MIPRO.2014.6859735

Justicia De La Torre, C., Sánchez, D., Blanco, I., & Martín-Bautista, M. J. (2018). Text Mining: Techniques, Applications, and Challenges. *International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems, 26*(4), 553-582. doi:10.1142/S0218488518500265

Kent, E. L. (2014). Text analytics – techniques, language and opportunity. *Business Information Review, 31*(1), 50-53. doi:10.1177/0266382114529837

Khan, W., Daud, A., Nasir, J. A., & Amjad, T. (2016). A survey on the state-of-the-art machine learning models in the context of NLP. *Kuwait Journal of Science, 43*(4), 95-113.

Kumar, P., & Wasan, S. K. (2010, 26-28 Feb. 2010). *Analysis of X-means and global k-means USING TUMOR classification.* Paper presented at the 2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE).

Kwok, J. T., Zhou, Z. H., & Xu, L. (2015). Machine learning. In *Springer Handbook of Computational Intelligence* (pp. 495-522).

Lee, J. H., Shin, J., & Realff, M. J. (2018). Machine learning: Overview of the recent progresses and implications for the process systems engineering field. *Computers and Chemical Engineering, 114*, 111-121. doi:10.1016/j.compchemeng.2017.10.008

Levy, Y., & Ellis, T. J. (2006). A Systems Approach to Conduct an Effective Literature Review in Support of IS Research. *Informing Science Journal, 9*, 181-212. doi:https://doi.org/10.28945/479

Louridas, P., & Ebert, C. (2016). Machine Learning. *IEEE Software, 33*(5), 110-115. doi:10.1109/MS.2016.114

Lv, H., & Tang, H. (2011, 22-23 Oct. 2011). *Machine Learning Methods and Their Application Research.* Paper presented at the 2011 2nd International Symposium on Intelligence Information Processing and Trusted Computing.

Manning, C. D., Raghavan, P., & Schütze, H. (2009). Scoring, term weighting and the vector space model. In An Introduction to Information Retrieval. Cambridge, England: Cambridge University Press.

Mishra, N., Agrawal, J., & KumarPatidar, A. (2012). Analysis of Different Similarity Measure Functions and Their Impacts on Shared Nearest Neighbor Clustering Approach. *International Journal of Computer Applications, 40*(16), 1-5. doi:10.5120/5061-7221

Miškuf, M., Michalik, P., & Zolotová, I. (2017). Data Mining in Cloud Usage Data with Matlab´s Statistics and Machine Learning Toolbox. *IEEE 15th International Symposium on Applied Machine Intelligence and Informatics*, 377-382.

Nakamura, C. M., Murphy, S. K., Christel, M. G., Stevens, S. M., & Zollman, D. A. (2016). Automated analysis of short responses in an interactive synthetic tutoring system for introductory physics. *Physical Review Physics Education Research, 12*(1). doi:10.1103/PhysRevPhysEducRes.12.010122

Omran, A. M. B., & Ab Aziz, M. J. (2013). Automatic essay grading system for short answers in English language. *Journal of Computer Science, 9*(10), 1369-1382. doi:10.3844/jcssp.2013.1369.1382

Ozgur, C., Colliau, T., Rogers, G., Hughes, Z., & Myer-Tyson, E. (2017). MatLab vs. Python vs. R. *Journal of Data Science 15*, 355-372.

Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing comprehension measured by multiple-choice and open-ended questions. *Can J Exp Psychol, 67*(3), 215-227. doi:10.1037/a0032918

Patil, S. K., & Shreyas, M. M. (2017). *A Comparative Study of Question Bank Classification based on Revised Bloom's Taxonomy using SVM and K-NN*. Paper presented at the 2017 2nd International Conference On Emerging Computation and Information Technologies (ICECIT).

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2008). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems, 24*(3), 45-77. doi:10.2753/MIS0742-1222240302

Pelleg, D., & Moore, A. (2000). *X-means: Extending K-means with Efficient Estimation of the Number of Clusters.* Paper presented at the ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning.

Perera, G. R., Perera, D. N., & Weerasinghe, A. R. (2016). *A dynamic semantic space modelling approach for short essay grading*.

Pérez-Delgado, M. L., Escuadra, J., & Antón, N. (2010) An Improved AntTree Algorithm for Document Clustering. In*: Vol. 79. Advances in Intelligent and Soft Computing* (pp. 481-488).

Phandi, P., Chai, K. M. A., & Ng, H. T. (2015). *Flexible domain adaptation for automated essay scoring using correlated linear regression.* Paper presented at the Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.

Pinckard, R. N., McMahan, C. A., Prihoda, T. J., Littlefield, J. H., & Jones, A. C. (2009). Short-answer examinations improve student performance in an oral and maxillofacial pathology course. *Journal of Dental Education, 73*(8), 950-961.

Pribadi, F. S., Permanasari, A. E., & Adji, T. B. (2018). Short answer scoring system using automatic reference answer generation and geometric average normalized-longest common subsequence (GAN-LCS). *Education and Information Technologies*. doi:10.1007/s10639-018-9745-z

Quah, J. T. S., Lim, L. R., Budi, H., & Lua, K. T. (2009). *Towards automated assessment of engineering assignments*.

Rahim, N. H. A. (2017). Fuzzy concepts compression using Principal Component Analysis with Singular Value Decomposition. *ARPN Journal of Engineering and Applied Sciences, 12*(2), 305-309.

Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing, 18*(1), 25-39. doi:https://doi.org/10.1016/j.asw.2012.10.004

Rangra, K., & Bansal, K. L. (2014). Comparative Study of Data Mining Tools. *International Journal of Advanced Research in Computer Science and Software Engineering, 4*(6), 216-223.

Rouhani, B. D., Mahrin, M. N. r., Nikpay, F., Ahmad, R. B., & Nikfard, P. (2015). A systematic literature review on Enterprise Architecture Implementation Methodologies. *Information and Software Technology, 62*, 1-20. doi:10.1016/j.infsof.2015.01.012

Roy, S., Narahari, Y., & Deshmukh, O. D. (2015) A perspective on computer assisted assessment techniques for short free-text answers. In*: Vol. 571* (pp. 96-109).

Sangodiah, A., Ahmad, R., & Ahmad, W. F. W. (2014). *A review in feature extraction approach in question classification using Support Vector Machine*. Paper presented at the 2014 IEEE International Conference on Control System, Computing and Engineering (ICCSCE 2014), Penang, Malaysia.

Shanie, T., Suprijadi, J., & Zulhanif. (2017). *Text grouping in patent analysis using adaptive K-means clustering algorithm.* Paper presented at the AIP Conference Proceedings.

Shehab, A., Elhoseny, M., & Hassanien, A. E. (2017). *A hybrid scheme for automated essay grading based on LVQ and NLP techniques*.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing, 20*, 53-76. doi:https://doi.org/10.1016/j.asw.2013.04.001

Srihari, S., Collins, J., Srihari, R., Srinivasan, H., Shetty, S., & Brutt-Griffler, J. (2008). Automatic scoring of short handwritten essays in reading comprehension tests. *Artificial Intelligence, 172*(2), 300-324. doi:https://doi.org/10.1016/j.artint.2007.06.005

Stanger-Hall, K. F. (2012). Multiple-choice exams: an obstacle for higher-level thinking in introductory science classes. *CBE Life Sci Educ, 11*(3), 294-306. doi:10.1187/cbe.11-11-0100

Swartz, S. M. (2010). Acceptance and Accuracy of Multiple Choice, Confidence-Level, and Essay Question Formats for Graduate Students. *Journal of Education for Business, 81*(4), 215-220. doi:10.3200/joeb.81.4.215-220

Talib, R., Hanify, M. K., Ayeshaz, S., & Fatimax, F. (2016). Text Mining: Techniques, Applications and Issues. *International Journal of Advanced Computer Science and Applications, 7*(11), 414-418. doi:http://dx.doi.org/10.14569/IJACSA.2016.071153

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly, 27*(3), 425-478.

Vijayarani, S., Ilamathi, J., & Nithya. (2015). *Preprocessing Techniques for Text Mining - An Overview.* Paper presented at the International Journal of Computer Science & Communication Networks.

Visvanathan, M., Srinivas, A. B., Lushington, G. H., & Smith, P. (2009). *Cluster validation: An integrative method for cluster analysis.* Paper presented at the Proceedings - 2009 IEEE International Conference on Bioinformatics and Biomedicine Workshops, BIBMW 2009.

Wang, H., Ma, C., & Zhou, L. (2009, 19-20 Dec. 2009). *A Brief Review of Machine Learning and Its Application.* Paper presented at the 2009 International Conference on Information Engineering and Computer Science.

Wang, H. C., Chang, C. Y., & Li, T. Y. (2008). Assessing creative problem-solving with automated text grading. *Computers and Education, 51*(4), 1450-1466. doi:10.1016/j.compedu.2008.01.006

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A Framework for Evaluation and Use of Automated Scoring. *Educational Measurement: Issues and Practice, 31*(1), 2-13. doi:10.1111/j.1745-3992.2011.00223.x

Witten, I. H., Don, K. J., Dewsnip, M., & Tablan, V. (2004). Text mining in a digital library. *International Journal on Digital Libraries, 4*(1), 56-59. doi:10.1007/s00799-003-0066-4

Wolska, M., Horbach, A., & Palmer, A. (2014) Computer-assisted scoring of short responses: The efficiency of a clustering-based approach in a real-life task. In*: Vol. 8686. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (pp. 298-310).

Wonowidjojo, G., Hartono, M. S., Frendy, Suhartono, D., & Asmani, A. B. (2016). *Automated essay scoring by combining syntactically enhanced latent semantic analysis and coreference resolution*.

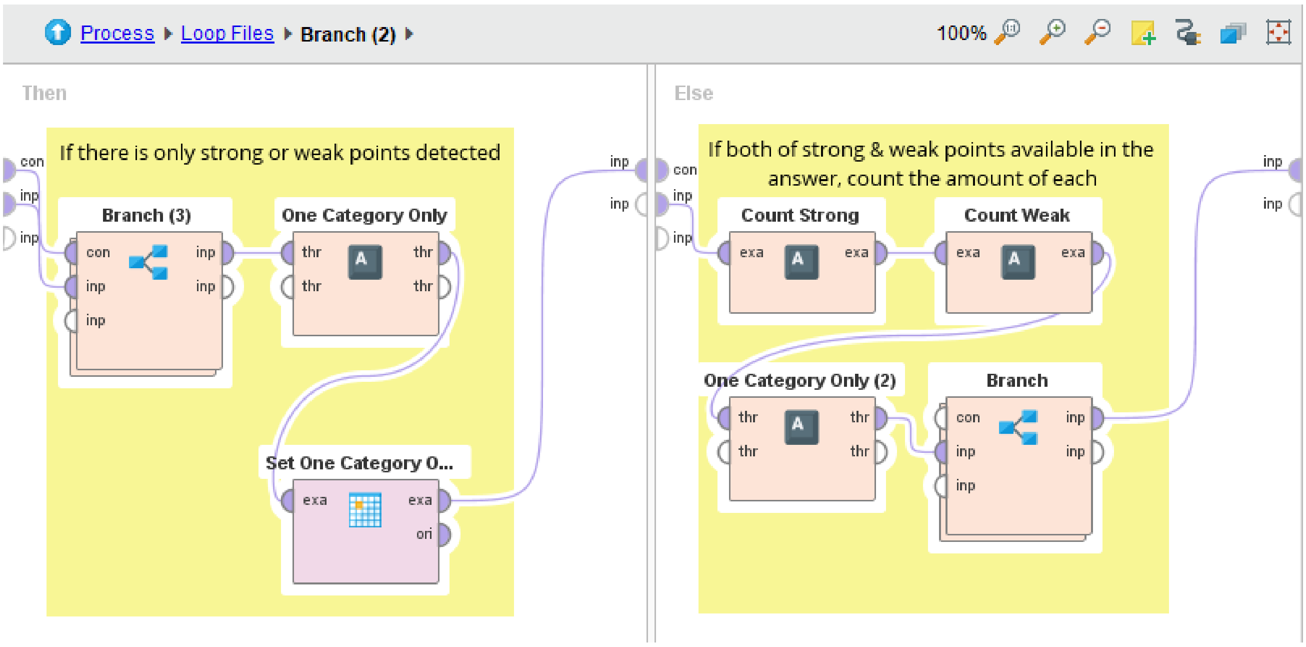
Wresch, W. (1993). The imminence of grading essays by computer—25 years later. *Computers and Composition, 10*(2), 45-58. doi:https://doi.org/10.1016/S8755-4615(05)80058-1

Zupanc, K., & Bosnić, Z. (2015). Advances in the field of automated essay evaluation. *Informatica (Slovenia), 39*(4), 383-395.

Zupanc, K., & Bosnić, Z. (2017). Automated essay evaluation with semantic analysis. *Knowledge-Based Systems, 120*, 118-132. doi:https://doi.org/10.1016/j.knosys.2017.01.006

APPENDIX A – FIGURES OF PROCESS

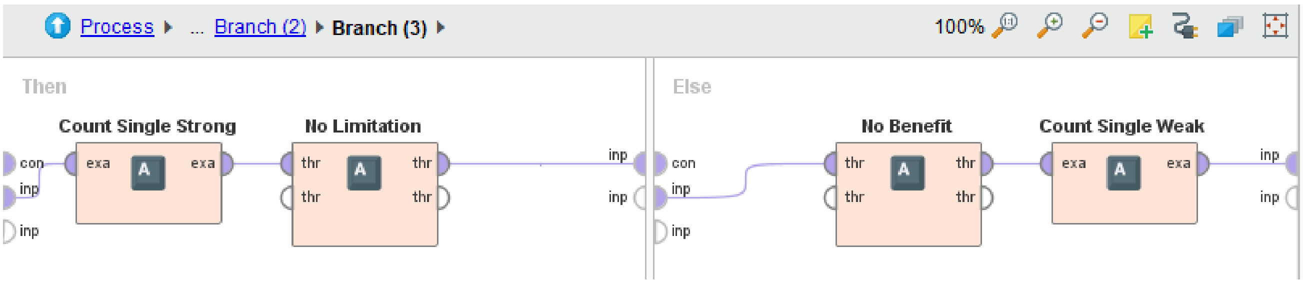
A.1. Process inside Branch (2) operator in Figure 17



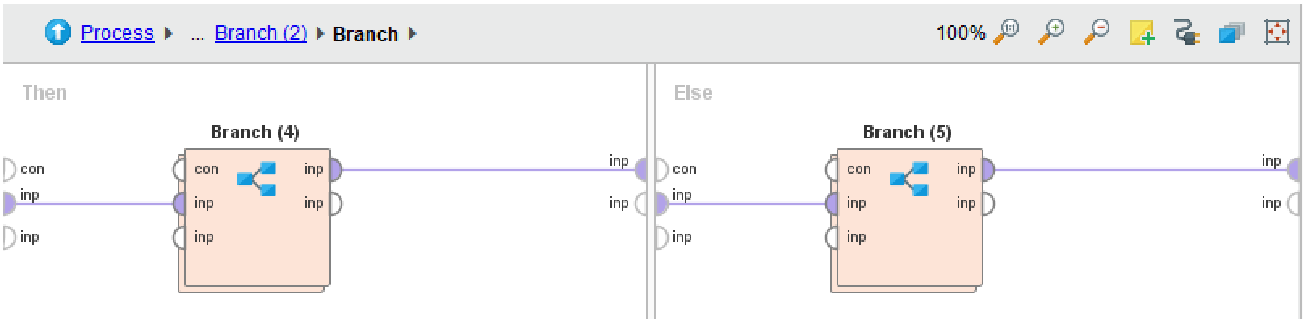
If there is only strong or weak points detected

If both of strong & weak points available in the answer, count the amount of each

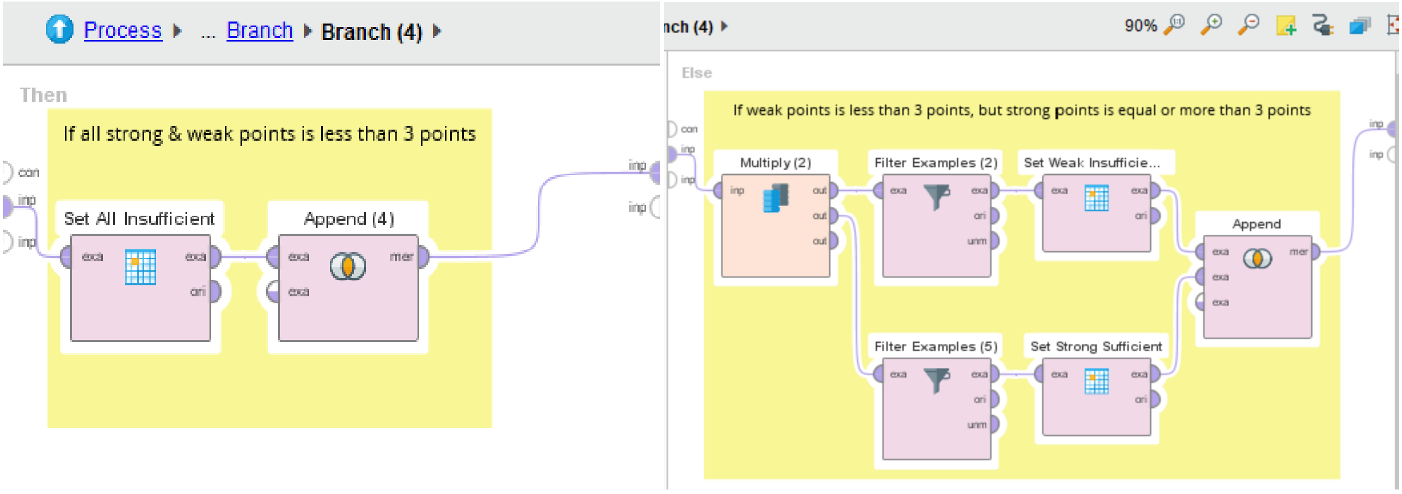
- Inside Branch (3)



- Inside Branch



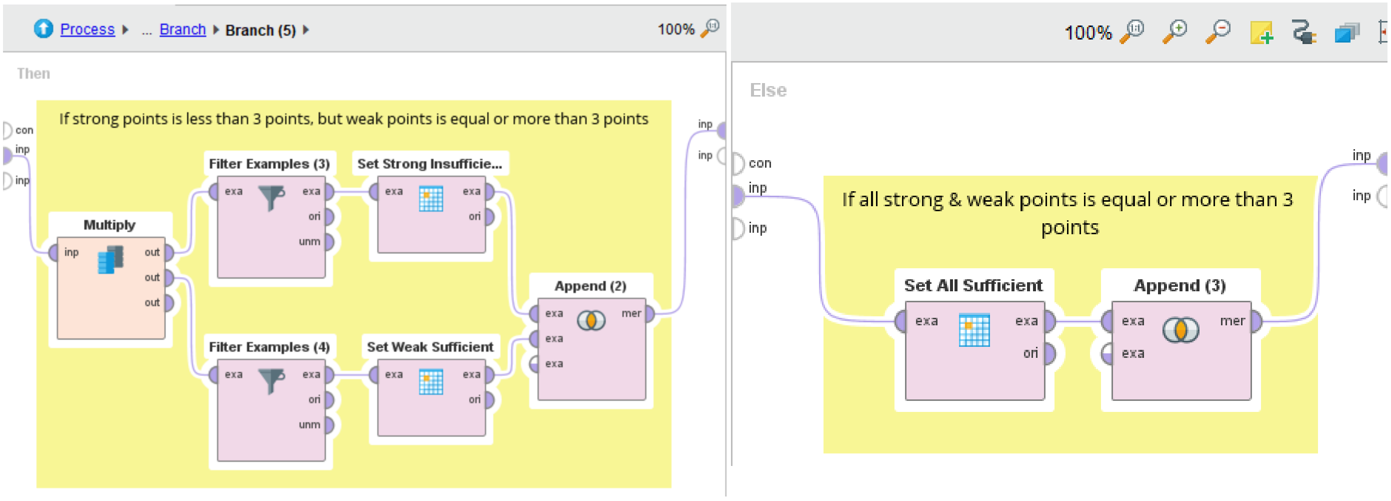
- Inside Branch (4)



If all strong & weak points is less than 3 points

If weak points is less than 3 points, but strong points is equal or more than 3 points

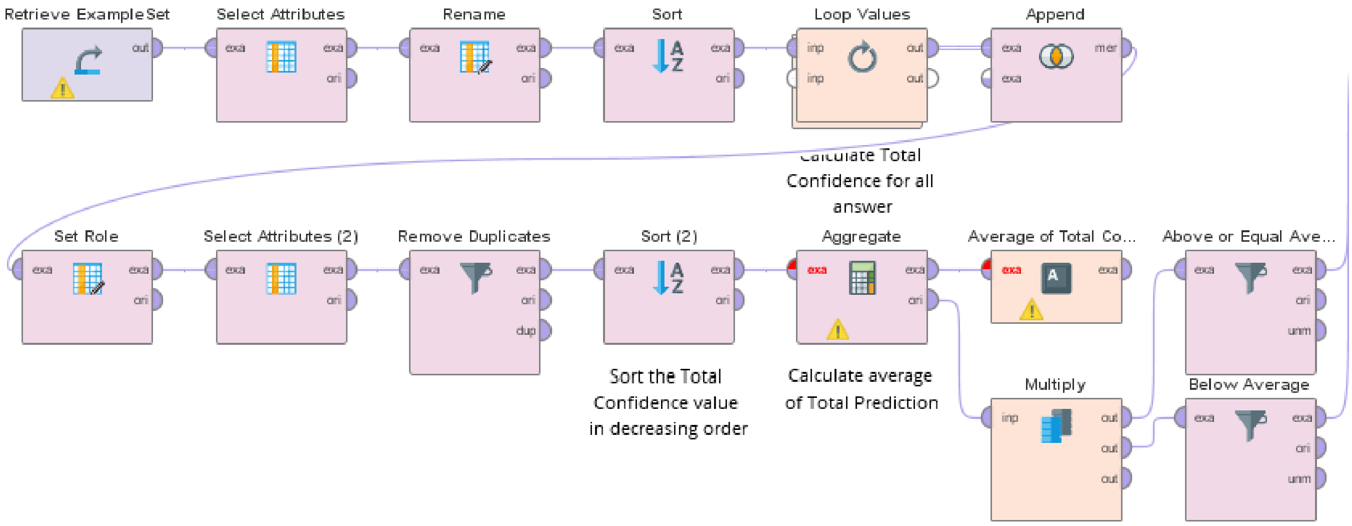
- Inside Branch (5)



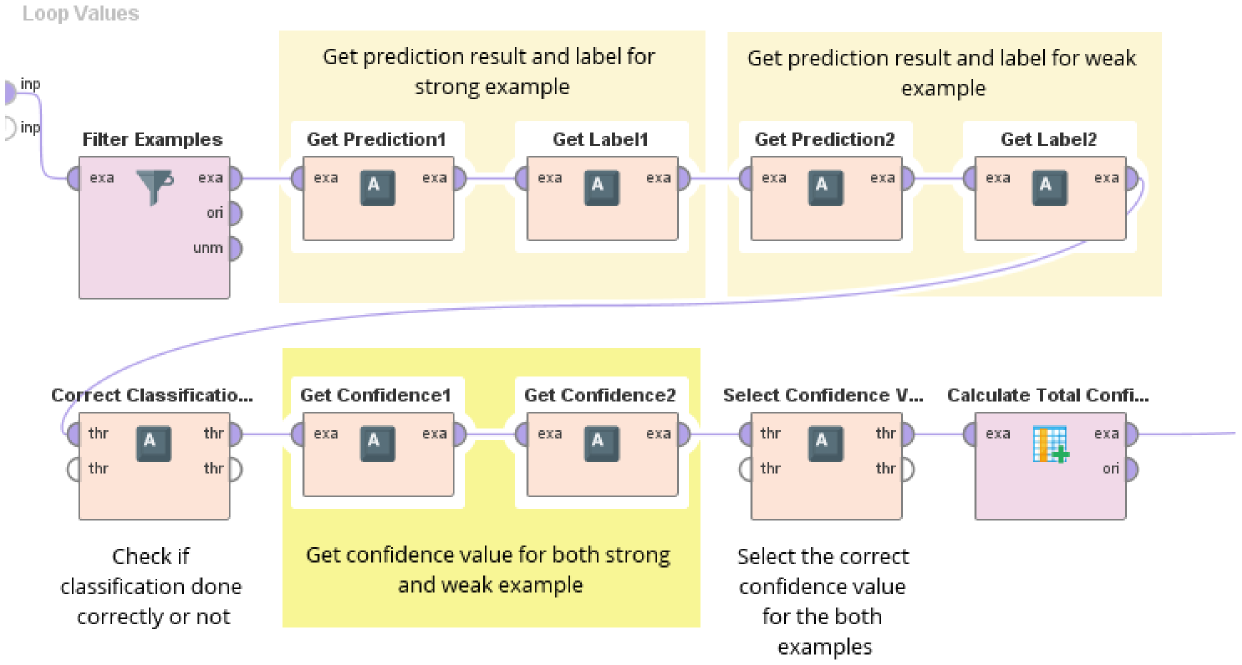
If strong points is less than 3 points, but weak points is equal or more than 3 points

If all strong & weak points is equal or more than 3 points

A.2. The whole process of ranking answers

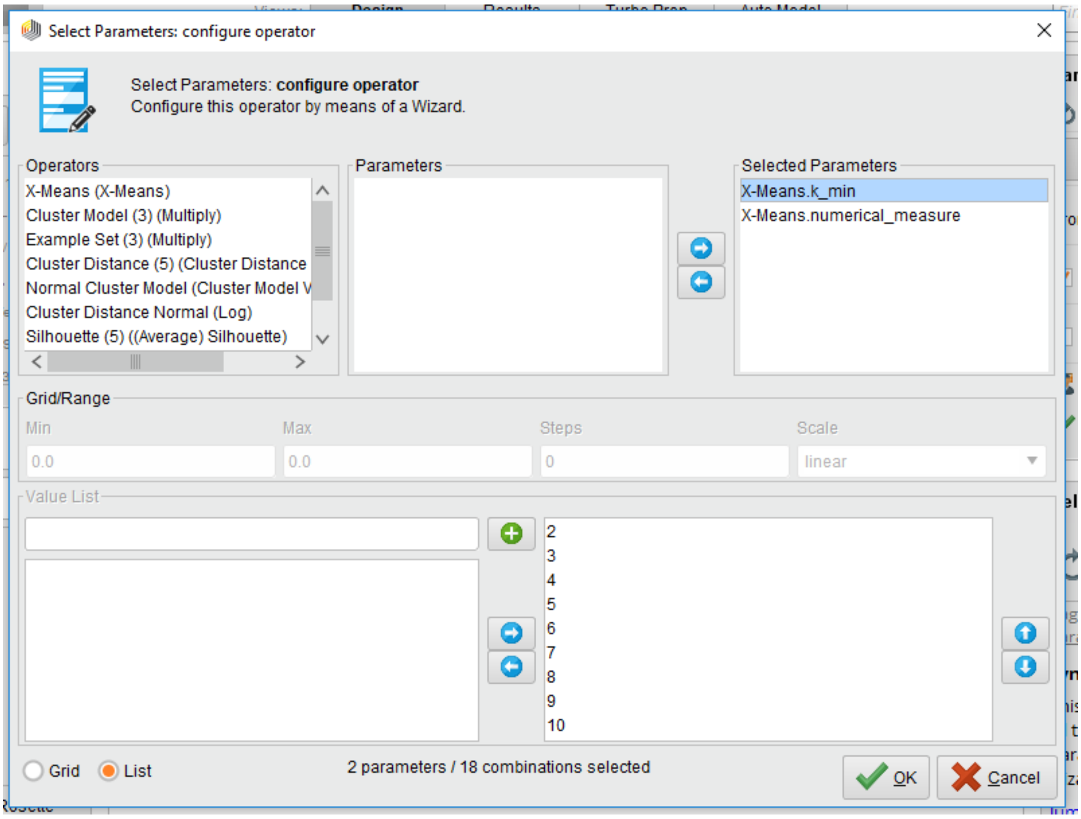


- Inside Loop Values Operator


Loop Values
inp
inp
Filter Examples
exa
exa
ori
unm
Get prediction result and label for strong example
Get Prediction1
Get Label1
Get prediction result and label for weak example
Get Prediction2
Get Label2
Correct Classificatio...
thr
Get Confidence1
Get Confidence2
Select Confidence V...
Calculate Total Confi...
Check if classification done correctly or not
Get confidence value for both strong and weak example
Select the correct confidence value for the both examples

A.3. The selection inside “Edit Parameter Settings”, one parameter of X-Means Loop Parameters operator

- The setting for k_min parameter



Select Parameters: configure operator

Select Parameters: **configure operator**
Configure this operator by means of a Wizard.

Operators
X-Means (X-Means)
Cluster Model (3) (Multiply)
Example Set (3) (Multiply)
Cluster Distance (5) (Cluster Distance
Normal Cluster Model (Cluster Model V
Cluster Distance Normal (Log)
Silhouette (5) ((Average) Silhouette)

Parameters

Selected Parameters
X-Means.k_min
X-Means.numerical_measure

Grid/Range
Min: 0.0
Max: 0.0
Steps: 0
Scale: linear

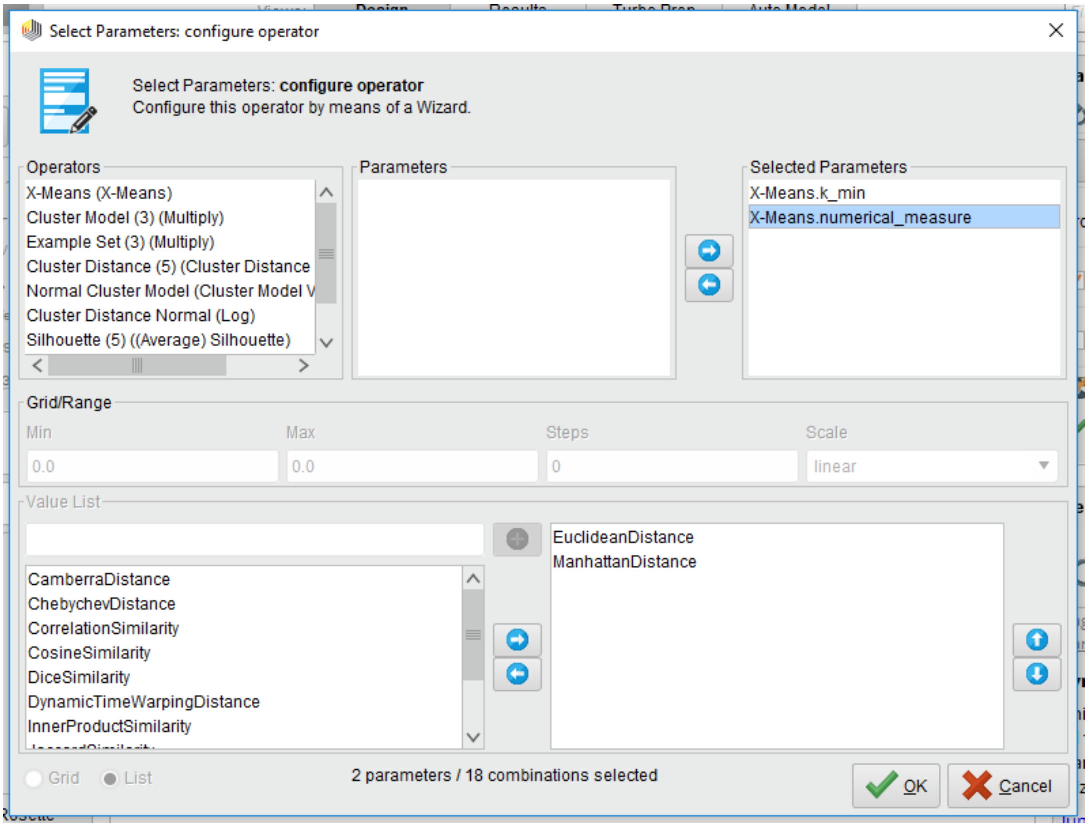
Value List
2
3
4
5
6
7
8
9
10

○ Grid ● List

2 parameters / 18 combinations selected

OK
Cancel

- The setting for numerical_measure parameter



Select Parameters: configure operator

Select Parameters: **configure operator**
Configure this operator by means of a Wizard.

Operators
X-Means (X-Means)
Cluster Model (3) (Multiply)
Example Set (3) (Multiply)
Cluster Distance (5) (Cluster Distance
Normal Cluster Model (Cluster Model V
Cluster Distance Normal (Log)
Silhouette (5) ((Average) Silhouette)

Parameters

Selected Parameters
X-Means.k_min
X-Means.numerical_measure

Grid/Range

Min	Max	Steps	Scale
0.0	0.0	0	linear

Value List

CamberraDistance
ChebychevDistance
CorrelationSimilarity
CosineSimilarity
DiceSimilarity
DynamicTimeWarpingDistance
InnerProductSimilarity

EuclideanDistance
ManhattanDistance

○ Grid ● List

2 parameters / 18 combinations selected

APPENDIX B – TABLE OF RESULT

B.1. Table Splitting Result for Question 2a of E-commerce 2018 exam.

No	Real Benefit	Real Limitation	Predicted Benefit	Predicted Limitation
1	2	2	4	4
2	2	2	2	2
3	2	2	2	2
4	2	2	0	2
5	2	2	2	2
6	2	2	2	2
7	2	2	2	2
8	6	2	1	3
9	2	1	0	1
10	2	2	4	4
11	2	2	1	1
12	2	2	1	2
13	2	2	4	4
14	4	3	5	2
15	2	2	2	2
16	2	2	3	2
17	2	2	2	2
18	6	6	7	6
19	2	3	1	3
20	2	2	1	1
21	4	1	4	2
22	2	2	5	2
23	2	2	2	2
24	2	2	2	2
25	2	2	2	2
26	2	2	4	5
27	2	2	2	2
28	2	2	2	2
29	2	2	1	1
30	2	2	3	3

B.2. Table Splitting Result for Question 2 of BCD4IT 2018 exam

No	Real Strong	Real Weak	Predicted Strong	Predicted Weak
1	3	3	3	3
2	3	3	3	4
3	4	4	4	4
4	3	3	3	3
5	3	3	4	4
6	3	3	3	3
7	3	3	3	3
8	3	3	3	3
9	4	4	4	5
10	3	3	3	3
11	3	3	3	3
12	3	3	3	3
13	3	3	4	3
14	3	3	3	3
15	3	3	3	3
16	3	3	4	4
17	3	3	6	6
18	3	3	3	3
19	3	3	3	7
20	3	3	3	3
21	3	3	3	3
22	3	3	3	3
23	3	3	3	3
24	3	3	3	3
25	3	5	3	5
26	3	3	3	3
27	3	3	6	6
28	3	3	3	7
29	3	3	4	3
30	3	3	3	3
31	3	3	3	3
32	3	3	3	7
33	3	3	3	3
34	3	3	3	4
35	3	3	3	3
36	3	3	3	3
37	3	3	3	3
38	3	3	3	3
39	3	3	3	3
40	3	3	3	3
41	3	3	3	3

42	3	3	3	3
43	3	3	3	3
44	3	3	4	3
45	3	3	3	4
46	3	3	3	3
47	3	3	6	10

B.3. Question 2a E-commerce 2018 Exam

Naïve Bayes		
	Accuracy	Correlation
Training Set	88.64%	0.774
Evaluation Set	94.44%	0.894

SVM		
	Accuracy	Correlation
Training Set	90.91%	0.818
Evaluation Set	88.89%	0.778

k-NN		
	Accuracy	Correlation
Training Set	93.18%	0.865
Evaluation Set	88.89%	0.798

B.4. Question 2 BCD4IT 2018 Exam

Naïve Bayes		
	Accuracy	Correlation
Training Set	81.82%	0.668
Evaluation Set	60.71%	0.247

SVM		
	Accuracy	Correlation
Training Set	92.42%	0.858
Evaluation Set	85.71%	0.714

k-NN		
	Accuracy	Correlation
Training Set	80.30%	0.606
Evaluation Set	82.14%	0.645

B.5. Classification Result of Training Data

label	prediction	conf(weak)	conf(strong)	file	score	problem
strong	strong	0.492	0.508	1	7	mentioned executive summary, not well-explained point
weak	weak	0.519	0.481			
strong	strong	0.468	0.532	2	10	
weak	weak	0.550	0.450			
strong	strong	0.468	0.532	3	10	
weak	weak	0.522	0.478			
weak	weak	0.500	0.500	4	8	
strong	weak	0.516	0.484			mentioned executive summary
strong	weak	0.509	0.491	5	6	first point is weakness
weak	weak	0.567	0.433			weak explanation
strong	strong	0.433	0.567	6	10	
weak	weak	0.564	0.436			
weak	weak	0.529	0.471	7	9	
strong	strong	0.441	0.559			mentioned executive summary
weak	weak	0.594	0.406	9	10	
strong	strong	0.430	0.570			
strong	weak	0.516	0.484	10	7	
weak	weak	0.539	0.461			second point is strength, third point is more like an advice
strong	strong	0.451	0.549	12	10	
weak	weak	0.534	0.466			
weak	weak	0.561	0.439	13	10	
strong	strong	0.474	0.526			
weak	weak	0.518	0.482	15	8	weak argumentation
strong	strong	0.431	0.569			
weak	weak	0.552	0.448	16	7	
strong	weak	0.505	0.495			mentioned executive summary, not well-explained point
strong	strong	0.490	0.510	17	9	third point is confusing
weak	weak	0.542	0.458			
strong	strong	0.423	0.577	18	10	
weak	weak	0.521	0.479			
strong	strong	0.483	0.517	19	7	unnecessary and some pointless explanation
weak	weak	0.541	0.459			
strong	strong	0.463	0.537	20	10	
weak	weak	0.529	0.471			
strong	strong	0.481	0.519	21	9	mentioned executive summary
weak	weak	0.555	0.445			
strong	strong	0.416	0.584	24	8	weak explanation
weak	weak	0.502	0.498			
strong	strong	0.484	0.516	25	10	

weak	weak	0.525	0.475			
strong	strong	0.361	0.639	27	8	
weak	weak	0.521	0.479			second weakness is not really a weakness
strong	strong	0.443	0.557	28	10	
weak	weak	0.535	0.465			
weak	weak	0.527	0.473	29	7	
strong	strong	0.497	0.503			second point is strange, third point is partially incorrect
strong	strong	0.436	0.564	30	8	first point is not very good
weak	weak	0.590	0.410			
weak	weak	0.515	0.485	32	10	
strong	strong	0.474	0.526			
weak	weak	0.544	0.456	36	8	missing decision tree is not problem
strong	strong	0.369	0.631			
weak	weak	0.569	0.431	37	9	
strong	strong	0.455	0.545			
weak	weak	0.565	0.435	38	10	
strong	strong	0.412	0.588			
strong	strong	0.449	0.551	40	8	the business case is not credible, some info are missing
weak	weak	0.533	0.467			
weak	weak	0.520	0.480	41	9	confusing explanation
strong	strong	0.479	0.521			
strong	strong	0.474	0.526	42	8	third strength
weak	weak	0.577	0.423			
strong	strong	0.451	0.549	43	10	
weak	weak	0.540	0.460			
strong	weak	0.521	0.479	47	8	second strength is confusing
weak	weak	0.518	0.482			

B.6. Classification Result of Testing (Evaluation) Data

label	prediction	conf(weak)	conf(strong)	file	score	problem
strong	strong	0.410	0.590	8	8	could use a bit more explanation
weak	weak	0.629	0.371			
strong	strong	0.342	0.658	11	10	
weak	weak	0.560	0.440			
strong	strong	0.355	0.645	14	10	
weak	weak	0.556	0.444			
strong	strong	0.332	0.668	22	9	
weak	weak	0.513	0.487			
strong	strong	0.408	0.592	23	9	

weak	weak	0.603	0.397			
strong	strong	0.364	0.636	26	6	very brief and no explanation
weak	strong	0.488	0.512			
strong	strong	0.444	0.556	31	10	
weak	strong	0.485	0.515			
strong	weak	0.521	0.479	33	7	are rather vague and are not clearly identified
weak	weak	0.632	0.368			
strong	strong	0.444	0.556	34	7	the third strength says that the business case is correct while the first weakness says that not everything is included; some of the explanations are rather brief or unconvincing
weak	weak	0.570	0.430			second point is strength, third point is more like an advice
strong	strong	0.477	0.523	35	10	
weak	weak	0.647	0.353			
strong	strong	0.434	0.566	39	9	
weak	weak	0.610	0.390			
strong	strong	0.491	0.509	44	10	
weak	weak	0.563	0.437			
strong	weak	0.539	0.461	45	10	
weak	weak	0.609	0.391			
strong	strong	0.427	0.573	46	10	

B.7. Ranking Results of Training Data

Above Average				Below Average			
No.	Sum of Confidence	File	Real Score	No.	Sum of Confidence	File	Real Score
1	1.175	36	8	1	1.074	21	9
2	1.164	9	10	2	1.066	20	10
3	1.160	27	8	3	1.058	19	7
4	1.154	30	8	4	1.058	5	6
5	1.153	38	10	5	1.054	3	10
6	1.131	6	10	6	1.052	17	9
7	1.114	37	9	7	1.047	16	7
8	1.103	42	8	8	1.041	32	10
9	1.098	18	10	9	1.041	25	10
10	1.092	28	10	10	1.041	41	9
11	1.089	43	10	11	1.030	29	7
12	1.088	7	9	12	1.027	1	7
13	1.087	13	10	13	1.023	10	7

14	1.087	15	8	14	0.997	47	8
15	1.086	24	8	15	0.984	4	8
16	1.084	40	8				
17	1.083	12	10				
18	1.082	2	10				

B.8. Ranking Results of Testing Data

Above Average				Below Average			
No.	Sum of Confidence	File	Real Score	No.	Sum of Confidence	File	Real Score
1	1.240	22	9	1	1.148	26	6
2	1.219	8	8	2	1.135	46	10
3	1.208	11	10	3	1.129	33	7
4	1.201	14	10	4	1.113	34	7
5	1.195	23	9	5	1.071	31	10
6	1.176	39	9	6	1.070	45	10
7	1.158	35	10	7	1.035	44	10

B.9. Sentiment Analysis

No	Answer Files	Length of Answer	Polarity	Polarity Confidence
1	1	206.0	positive	0.699
2	10	367.0	positive	0.924
3	11	221.0	negative	0.499
4	12	351.0	neutral	0.654
5	13	406.0	positive	0.986
6	14	251.0	neutral	0.561
7	15	100.0	positive	0.976
8	16	201.0	neutral	0.848
9	17	60.0	positive	0.888
10	18	191.0	negative	0.812
11	19	194.0	negative	0.922
12	2	381.0	positive	0.992
13	20	260.0	neutral	0.509
14	21	203.0	positive	0.547
15	22	169.0	positive	0.541
16	23	272.0	positive	0.505
17	24	210.0	neutral	0.731
18	25	450.0	positive	0.693
19	26	96.0	neutral	0.577
20	27	334.0	positive	0.980
21	28	113.0	negative	0.363
22	29	186.0	neutral	0.543
23	3	182.0	positive	0.689
24	30	119.0	negative	0.950

25	31	539.0	neutral	0.539
26	32	312.0	neutral	0.917
27	33	312.0	positive	0.736
28	34	106.0	positive	0.642
29	35	261.0	positive	0.990
30	36	317.0	negative	0.502
31	37	254.0	positive	0.574
32	38	221.0	neutral	0.923
33	39	167.0	neutral	0.876
34	4	95.0	positive	0.622
35	40	337.0	neutral	0.820
36	41	210.0	positive	0.766
37	42	327.0	positive	0.954
38	43	202.0	neutral	0.382
39	44	280.0	positive	0.883
40	45	254.0	neutral	0.741
41	46	137.0	neutral	0.578
42	47	115.0	positive	0.906
43	5	186.0	positive	0.806
44	6	339.0	neutral	0.970
45	7	158.0	positive	0.881
46	8	150.0	positive	0.970
47	9	177.0	neutral	0.511

APPENDIX C – QUESTIONNAIRE DESIGN

Questionnaire for Evaluation Workshop

Applying Text Mining and Machine Learning to Build Methods for Automated Grading

My name is Febriya Hotriati Psalmerosi. I am a second year Master student in Business Information Technology at the University of Twente. Currently, I am working in my thesis about building methods by applying text mining and machine learning for automated grading.

This form is used to get the opinion from the teachers about proposed methods. There are 6 sections and 17 statements in total. Each section consists of 2-4 statements in which the participants should choose their preference about the statements. Please circle the number that represents your level of agreement with the statement. The result will be analyzed as the Evaluation result of my Master thesis.

Thank you.

Statements	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
Performance Expectancy					
1. I would find the proposed method useful in grading an answer.	1	2	3	4	5
2. Using the proposed method enables me to grade more quickly.	1	2	3	4	5
3. Using the proposed method increases my productivity.	1	2	3	4	5
Effort Expectancy					
4. My interaction with the proposed method would be clear and understandable.	1	2	3	4	5
5. I would find the proposed method easy to use.	1	2	3	4	5
6. It would be easy for me to become skillful at understanding the proposed method.	1	2	3	4	5
Facilitating Conditions					
7. I have the resources necessary to use the proposed method.	1	2	3	4	5
8. I have the knowledge necessary to use the proposed method.	1	2	3	4	5
9. The proposed method is not compatible	1	2	3	4	5

with other systems I use.

Self-efficacy

10.	I could grade using the proposed method if there was no one around to tell me what to do as I go.	1	2	3	4	5
11.	I could grade using the proposed method if I could call someone for help if I got stuck.	1	2	3	4	5

Attitude toward using technology

12.	Grading answer with the proposed method is a good idea.	1	2	3	4	5
13.	The proposed method makes grading more interesting.	1	2	3	4	5
14.	Grading with the proposed method would be fun.	1	2	3	4	5

Behavioral Intention to Use

15.	I intend to use the proposed method in the future to help me grading student's answer.	1	2	3	4	5
16.	I predict I would use the proposed method in the future to help me grading student's answer.	1	2	3	4	5
17.	I plan to use the proposed method in the future to help me grading student's answer.	1	2	3	4	5

Additional Feedback

Please fill in any additional feedback(s) regarding the proposed method.