Student Thesis

Level: Master

Exploration of an Automated Motivation Letter Scoring System to Emulate Human Judgement

Author: Lorenna Munnecom, Miguel Chaves de Lemos Pacheco
Supervisor: Asif M Huq and Kenneth Carling
Examiner: Moudud Alam
Subject/main field of study: Micro Data Analysis
Course code: MI4001
Credits: 30
Date of examination: June 8, 2020

Dalarna University – SE-791 88 Falun – Phone +4623-77 8

Contents

I. Abstract

As the popularity of the master's in data science at Dalarna University increases, so does the number of applicants. The aim of this thesis was to explore different approaches to provide an automated motivation letter scoring system which could emulate the human judgement and automate the process of candidate selection. Several steps such as image processing and text processing were required to enable the authors to retrieve numerous features which could lead to the identification of the factors graded by the program managers. Grammatical based features and Advanced textual features were extracted from the motivation letters followed by the application of Topic Modelling methods to extract the probability of each topics occurring within a motivation letter. Furthermore, correlation analysis was applied to quantify the association between the features and the different factors graded by the program managers, followed by Ordinal Logistic Regression and Random Forest to build models with the most impactful variables. Finally, Naïve Bayes Algorithm, Random Forest and Support Vector Machine were used, first for classification and then for prediction purposes. These results were not promising as the factors were not accurately identified. Nevertheless, the authors suspected that the factors may be strongly related to the highlight of specific topics within a motivation letter which can lead to further research.

Key Words

Natural Language Processing, Machine Learning, Supervised Learning, Unsupervised Learning, Automation, Feature Extraction, Image Processing, Text Processing, Text Exploration, Motivation Letter, Dalarna University, Student Application, Topic Modelling, Business Intelligence, Data Science

II. List of tables

III. List of figures

IV. List of Equation

V. Abbreviations

ATS: Automatic Text Scoring

TOEFL: Test of English as a Foreign Language

GMAT: Graduate Management Admissions Test

GRE: Graduate Record Examination

OLR: Ordinal Logistic Regression

DS: Data Science

BI: Business Intelligence

NLP: Natural Language Processing

NLTK: Natural Language Toolkit

POS: Part-of-Speech

SMOG: Simple Measure of Gobbledygook

BOW: Bag of Words

TF-IDF: Term Frequency-Inverse Document Frequency

LSA: Latent semantic analysis

LDA: Latent Dirichlet Allocation

STM: Structural Topic Model

NBA: Naive Bayes Algorithm

RF: Random Forest

SVM: Support Vector Machine

1. Introduction

The re-shaping of the former master's program in Micro Data Analysis in to a one-year Master's in Business Intelligence and to a two-year Master's in Data Science coming into effect as of the fall 2019 has drastically increased the number of applicants at Dalarna University. For each intake there are about 1000 applicants competing for 25 slots. Consequently, there is a need for a selection instrument. For the intake of the fall of 2020, applicants could have voluntarily submitted a motivation letter, which was used to rank the applicants. The masters received 365 motivation letters. From the 365 letters, 221 were manually reviewed by two program managers who scored them according to 7 factors. The others were not considered, because the students did not meet the requirements. Each program manager read a given number of motivation letters and attributed a score.

A motivation letter is a letter of introduction which accompanies the student application at Dalarna University. The university requires that a motivation letter should not have more than 3000 characters, including spaces. It should also state the academic and professional history of the applicant, with relevance to the program, and explain how the student pretends to benefit from the program on her/his future career (Study at Dalarna Univeristy, 2020).

This thesis explores ways to assist on the automation of a motivation letter scoring system which emulates human judgement. Nonetheless most research papers focus on the automated scoring of essays. According to Cambridge online dictionary an essay is "a short piece of writing on a particular subject, especially one done by student as part of the work for a course".

Reading and assessing hundreds of essays is a time-consuming task and comprises of issues such as objectivity and accountability. A way to face the issues from manual scoring is to have an Automated System for Essay Evaluation (AEE). AEE can be described as the process of assessing and scoring essays using algorithms (Burstein, 2003). AEE is a field of multiple disciplines such as computer science, linguistics and psychology. The first automated essay scoring (AES) was proposed and designed by Ellis Batten Page and his colleagues (Page, 1966). The main motivation for the creation of this software was that it was believed that the more students wrote, the better they would

write. Therefore, attributing more essays as homework would not be a burden to the professors.

Throughout the research, several papers were found to focus on the quality and assessment of student essays. If an essay is used for training purposes, they are named low-stake essays, while an essay that may have serious consequences for the students are called high-stake essays (Smolentzov, 2013). An example of high-stake essay is one of TOEFL test sections, as the test objective is to evaluate people's English language skills to certify their ability to follow courses in English at the university. Given the definition, motivation letters are comparable to high-stake essays.

Automated Text Scoring (ATS) is known to be a Natural Language Processing application generally used in the educational field to automatically analyse student's response to questions or to analyse student's essays (Madnani & Cahill, 2018). Natural Language Processing is how machines analyse, understand and derive meaning from human language in a smart and useful way (Lopez & Kalita, 2017). Three main stakeholders were identified in the ATS process, first the program managers, by saving them time and because of their participation on the calibration of the system to find the most suitable students for the program. Second, the students, as a proper fit will lead to a higher student satisfaction with the programme. Last, Dalarna University students that fit better will be more likely to promote the university and the program. Furthermore, the university's prestige and ability to sell itself among future students will be enhanced.

Although many papers gave emphasis to essays, it is believed that many of the procedures and algorithms used may be applied to this thesis research, which aims to explore ways of assisting the automation of motivation letters scoring system to emulate human judgement to provide with reliable, accountable and objective scoring of the motivation letters submitted by the applicants. By focusing on such topic, it is possible to avoid problems such as theme diversity, type of language, ability to focus on the semantic structure, the context and objective of the motivation letter.

Reading and assessing hundreds of Motivation Letter is a time-consuming task and comprises of issues such as objectivity and accountability. As the popularity of the master's in data science at Dalarna University increases, so does the number of applicants therefore there is a need for selection instruments. The aim of this thesis was to explore different approaches to provide an automated motivation letter scoring system

which could emulate the human judgement and automate the process of candidate selection.

The program managers did not disclose the factors on which a candidate was evaluated therefore the research question captured was:

- Which features can be identified and how do they influence a Motivation Letter scoring?

To answer the research question, several steps such as data collection, data pre-processing, image processing and text processing were required to enable the authors to retrieve numerous features which could lead to the identification of the factors graded by the program managers. Grammatical based features and Advanced textual features were extracted from the motivation letters followed by the application of Topic Modelling methods to extract the probability of each topics occurring within a motivation letter. Furthermore, correlation analysis was applied to quantify the association between the features and the different factors graded by the program managers, followed by Ordinal Logistic Regression and Random Forest to build models with the most impactful variables. Finally, Naïve Bayes Algorithm, Random Forest and Support Vector Machine were used, first for classification and then for prediction purposes. Results, conclusion and discussion following the application of these methods are further explained in the next sections.

2. Start-of-the-art Review

Automated Text Scoring (ATS) had its creation with Project Essay Grade in 1966 by high school teacher Ellis Page (Page, 1966) and is one of the first automated scoring systems, which would predict the score using linear regression over vectors of text features that were thought to represent the quality of the writing. The system would allow teachers to avoid hours of manually grading student essays. At that time, the necessary resources needed to scale this idea were simply not available. Furthermore, the idea of having humans replaced by systems was not welcomed by society (Shermis et al., 2013). ATS can be defined as the statistical and natural language process techniques used to automatically score a text on a marking scale. (Alikaniotis et al., 2016).

Nowadays, the name that replaced ATS is Automated Essay Scoring (AES) or Automated Essay Evaluation (AEE) which has been defined as the task of using computer technology

to score written text (Ke & Ng, 2019). The main reasons that have contributed to the interest of AES are: cost, accountability, standards and technology (Harshanthi, 2019).

With the expansion of the internet, AES became an important technological tool in education. The AES is used to score essays alongside human graders and are tools used by high-stake essays, such as the Test of English as a Foreign Language (TOEFL), Graduate Management Admissions Test (GMAT) and Graduate Record Examination (GRE).

There are many systems on the market, and mostly have a commercial purpose. In (Zupanc, 2018) 20 state-of-the-art systems where compared by using attributes like style, content, semantic and methods used for extraction of features. In the end, it mentions the prediction models used by each system to score the essays (Appendix 1).

Before applying models, measurable features first need to be successfully identified and extracted. Many studies demonstrated that Coh-Metrix is a powerful tool that analyses texts on various linguistic features. Coh Metrix is a computer tool developed by Arthuc C. Graesser and Danielle S. McNamara in 2004. "The tool analyses texts on over 200 measures of cohesion, language and readability. Its modules use lexicons, part-of-speech classifiers, syntactic parsers, templates, corpora, latent semantic analysis and other components that are widely used in computational linguistic" (Arthur et al., 2004).

A supervised machine learning approach for AES was applied to each essay that was represented by a feature vector (Östling et al., 2013). The authors have modelled several features. The simple features evaluate the text length, average word length, lexical diversity and speech distribution. The text length refers to the number of words in a document while the average word length refers to character count divided by the number of words. Lexical Diversity also called lexical richness is defined as the ratio of different unique words in the document. The second feature called *Corpus-induced feature* compares student essays to statistics gathered from different text sources. Which means, that vocabulary and grammar are being compared to different online text sources to compute the average cross-entropy, also defined as the probability of events occurring. Finally, the language error feature involves a simple spell check.

A neural network architecture model was developed in order to capture the local linguistic environment of each word (Collobert & Weston, 2008). Usage information are captured

by the linear order of the words in a sentence, which means that under-informative words such as auxiliary verb will not have influence on the essay score, whereas informative words will heavily influence the score.

An extended neural network architecture was applied by (Alikaniotis et al., 2016). To capture their so-called score-specific word embeddings the model was extended so it would also capture how each word contributes to the overall score of the essay, thus a linear unit in the output layer was added to predict the essay score with a linear regression (Alikaniotis et al., 2016).

An Ordinal Logistic Regression (OLR) which uses a length-only classifier and Naïve Bayes as baselines was also presented (Woods et al. 2017). The length classifier addresses the correlation between essay length and score using a logistic regression classifier with the character length of the essay as predictor. The Naïve Bayes model uses binary features indicating the presence of word, character, and part of speech n-grams. While the two baselines are not enough, the ordering of information was not being captured and the categorical nature of the score would suffer with the linear regression approach, the ordinal model accounts for both. The feature used in OLR model includes essay length, count of words, character and POS n-grams, binary indicator of the same n-grams (Woods et al,. 2017).

On a similar topic, (Harshanthi, 2019) wrote about "Automated Essay Evaluation using Natural Language Processing and Machine Learning". The paper focus on the inter-rate reliability and performance of AEE. The paper takes a traditional approach where the features used to evaluate the essays are lexical diversity, sentence count, word frequency, word count, average length, structure and organization of an essay. (Harshanthi, 2019)used 3 algorithms to compare against the human scorer. The algorithms chosen were Linear Regression, Random Forest Regression and Support Vector Regression. As this paper deals with motivation letters, which will have an impact on the class and composition of the Master class, it is important to make sure that different human characteristics are identified and selected so that the future group of students may be, from an educational point of view, diverse enough.

According to (Zupanc, 2018), text semantics are not properly explored which leads to be one of the main weakness of the existing systems. Evaluation of essays in (Zupanc, 2018) is approached according to coherence and semantic error detection. Also, it is proposed

the Automated Error Detection (AED) system where the essay semantics are analysed from the perspective of essay consistency. According to the paper, the proposed system SAGE achieves higher grading accuracy when faced against other systems.

3. Methodology

3.1. Data Collection

Sweden offers the convenience for candidates to apply for a program electronically via Universityadmissions.se. The deadline for the submission of application for the Data Science and Business Intelligence master program is before 15th of January 2020 for the Autumn semester and 15th of August for the Spring semester.

To complete their electronic application, candidates must submit documentation to demonstrate their eligibility to the program they have applied for. To be eligible, candidates must meet certain general and specific entry requirements. The candidate is required to have a bachelor's degree with a relevant field of study which is related to computer science, computer information system and data science. The degree should come from an international recognized university and the candidate should demonstrate proficiency in English by taking an internationally recognized test.

Besides, the letter of motivation is one of the selection criteria, the letter should not exceed 3000 characters, including spaces and it should state the academic and professional history of the applicant, with relevance to the programme. It is also requested a statement on how the programme will benefit the future career of the applicant.

The candidate uploads the documentation to the University Admissions website. The University Admissions is managed by the Swedish Council for Higher Education, whose main mission is to review applications to see if students meet the general entry requirements. The applications are then checked by different institutions and once the pre-requisites were approved by the admission office, the candidate documentation is sent to the university. As the program managers are notified of potential candidates for the programs by the University Admission, they are required to retrieve the motivation letters, read them and score them for each eligible candidate (Dalarna University, 2020).

3.2. Data Description

A total of 139 images containing candidate motivation letters where extracted for the BI program and 192 images where extracted for the Data Science program for the eligible

applicants that had submitted a motivation letter. The motivation letters were extracted from the system in the form of images in .png format.

Additionally, a scoring file in excel format for BI and DS application was provided by the program managers. The scoring file contained information related to the candidate such as person-number, first name, family name, educational background and the program priority rank selected by the candidate. The most significant attribute and metrics from the scoring file are the image IDs assigned to each motivation letter, an overall score between 0 and 50 and the seven factors with grade level of 0, being the lowest followed by 1 and 2 being the highest.

Although the seven factors have an influence on the overall score, the sum of these factors does not represent the final score. Since the seven factors somehow influence the final score it was assumed there would be a clear linearity between the sum of factors and the score. Thus, our first hypothesis implied condensing the 7 factors as one-dimension problem which means that the lowest sum of factors would then lead to the lowest overall score while the highest sum of factors would lead to a highest overall score. As seen in figure 1, some correlation seemed evident but other questions arose. For example, it is possible to see that, there were applicants who had the same total sum of factors but were attributed different scores.

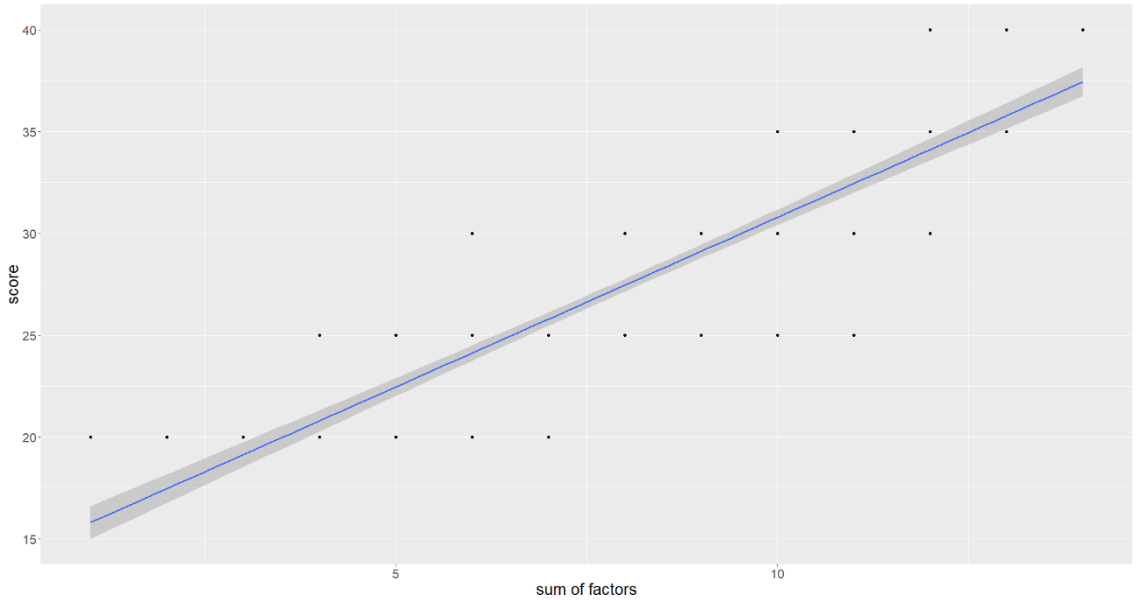


Figure 1 Scatter plot showing the relation between the sum of factors and overall grade

These observations guided us to other hypothesis, such as the discrepancies between the scoring process between the two program managers or that the different factors are weighted differently. Since it is not known which program manager scored which cover letter no further investigation could take place. For the second hypothesis, the research was carried in a partial way which means that one factor at the time was analysed, hence assuming independence between factors.

3.3. Data Processing

In order to explore the data several steps were necessary beforehand. Python is known to have a large community of people dealing with vision and image processing with quite interesting libraries related to vision thus the Image Processing was done in Python while Data Exploration and Data Analysis was further exploited in R.

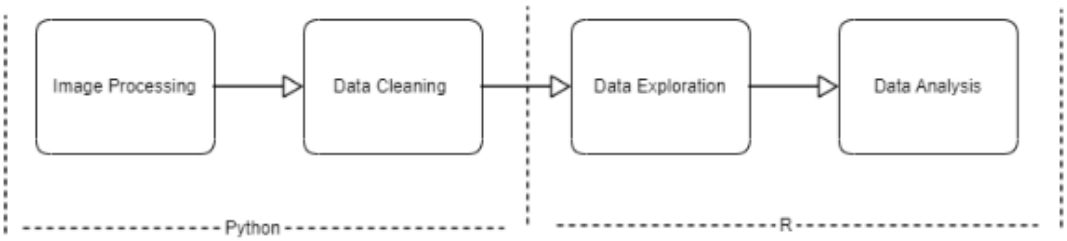


Figure 2 Data Processing Systems

3.3.1 Image Processing

331 images were provided by the program managers thus authors expected 331 motivation letters. Nevertheless, some motivation letters exceeded one page thus they were provided as different image files These image files were named identically with the only exception that the first page of the motivation letter would hold the image ID and a last character being ‘a’ while the second page of the motivation letter would hold the image ID and a last character being ‘b’ and so on.

A Python script was developed to build a dictionary to extract the pagination based on the last character of the naming convention and to group images with identical image IDs. This process returned the actual number of motivation letters to 225, 107 motivation letters for BI program and 118 motivation letters for DS program.

These images were then passed through the Python-Tesseract which is an optical character recognition (OCR) tool for Python. It is known to be the most popular and qualitative OCR-library. It recognises and read text embedded in images by finding templates in pixels, letters, words and sentences (Hoffstaetter, 2020). Figure 3 provides a visualisation of the step taken from the original image to text document.

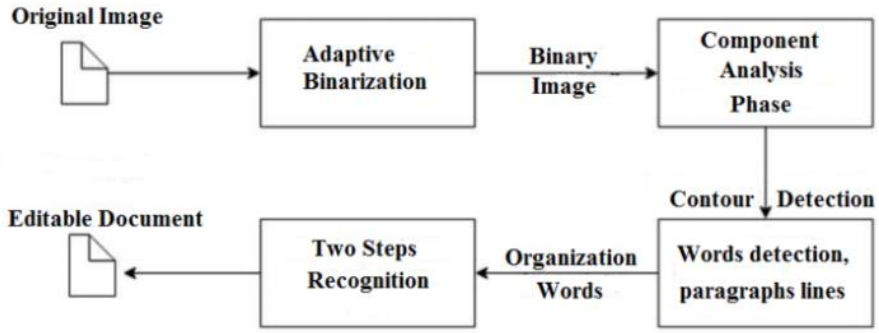


Figure 3 Tesseract OCR process from Boiangiu, 2016

The major issues faced by the authors were the images with low quality, difficulties to capture the main corpus of the motivation letter due to headers and bottoms of the text and some whitespaces which were not identified by Tesseract. The result demonstrated that Tesseract worked best when the files had proper dimensions with no background colour and when the parameters were adjusted. As Tesseract printed the text from the images, all converted images were saved and imported into a data frame to ease further data manipulation (Boiangiu, 2016).

3.3.2 Data Pre-Processing

The scoring file was imported in Python and transformed in a data frame. Among relevant information contained in the data frame, the ones with less relevance such as candidate personal information were excluded. The scoring data frame was then joined with the motivation letter data frame.

Figure 4 is the initial data frame which served as basis to extract additional information such as the program the candidate applied for, the diploma the candidate holds and the gender of the candidate.

	cover_letter	ID_image	prio	merit_rating	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
1	Letter of MotivationI am writing this letter to express my int...	BI1	4	25	2	2	2	0	0	0	1
2	Dear Sir/ Madam,I am writing to inform you of my interest i...	BI101	1	25	2	1	2	1	1	1	1
3	Motivation Letter Dear Sir/Madam,I am writing this letter wi...	BI102	2	30	2	2	2	1	1	1	1
4	Statement of PurposeIt is my absolute pleasure to express ...	BI103	4	20	1	0	0	1	0	1	0
5	STATEMENT OF PURPOSE Fransisco Kibasal would describe ...	BI104	3	20	2	1	0	1	1	1	1
6	MOTIVATION LETTERDhiraj Kumar 20, Soubhagya Nagar Bh...	BI105	2	30	2	2	0	1	2	1	2
7	Dalarna University/Hogskolan Dalarna Hodgskolegatan 2 79...	BI111	1	25	2	1	1	0	1	0	1

Figure 4 Data frame for cleaned data set

All rows with a score of 0 were excluded from the data frame as no motivation letter was identified. Data quality issues were identified on 3 motivation letters. These motivation letter had an overall score of 0 although the individual factors were graded. After checking with the program managers, their overall score was amended. A fourth motivation letter was identified and had not been scored since it didn't meet the requirement thus it was excluded, which reduced the dataset to 224 observations.

Besides, to respect the General Data Protection Regulation (GDPR) the identities of the applicants were anonymised. Candidates data are only available to the university admission office, the program managers and the authors of the thesis.

4. Natural Language Processing Modelling

As the objective was to identify the 7 factors that were graded in the motivation letter. The 7 factors were not disclosed by the program managers and the reason was because the authors were required to identify these through their research. To identify these factors, the authors had to use text cleaning methods for Natural Language Processing (NLP). Once the text was pre-processed, grammatical based features and advanced textual features were extracted. Grammatical based features denote to various kinds of information about grammatical entities whereas advanced textual features refer to various methods with advanced capabilities. Once the features were extracted, unsupervised learning methods for topic modelling was used and a supervised learning approach was applied on the features.

4.1 Feature extractions

4.1.1 Text Pre-Processing

Before diving into feature extraction, the following steps became imperative to mine actionable insights from the text. Python has a package for NLP called Natural Language Toolkit (NLTK) known as the most powerful NLP library as it contains packages to make machines understand human language.

After converting the images into text documents, the first text pre-processing step involved tokenizing the text of the motivation letters. “Tokenization is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words or terms” (Singh, 2019). The next step involved removing stop words, punctuation and transforming each term into lower case. Stop words are commonly word such as: “the”, “and”. The process of converting the words into its root word, known as stemming and lemmatization, was not applied as this resulted in words loss thus loss of text context.

4.1.2 Grammatical Based Feature

Grounded on the literature review, several features from the text where extracted. Basic grammar features as word count, character count, stop words count, syllable count, grammar error count, sentence count, difficult words count, misspelled words rate, and lexical diversity where extracted from the pre-processed motivation letter.

Additionally, semantic and stylistic features from text were extracted. The semantic features refer to a sequence of part-of-speech tagging (POS tagging). POS tagging allows to identify which words act as nouns, pronouns, verbs, adverbs and so on. The stylistic feature is a count of the tag occurrence within the corpus.

Besides, readability scores of texts were extracted. “Readability is the ease with which a reader can understand a written text. In natural language, the readability of text depends on its content (the complexity of its vocabulary and syntax) and its presentation (such as typographic aspect like font size, line height and line length)” (Holtzscher, 2019). For exploration purpose more than one metric was retrieved. The popular metrics calculated

were the Flesch-Kincaid Readability Score, the Flesch-Kincaid Reading Ease, the SMOG index and the Gunning FOG index.

The Flesch Formulas is an approach to assess the grade-level of the reader. It only uses two variables for readability, being word length and sentence length. The Gunning FOG index is comparing syllables and sentence lengths and the SMOG index counts the number of sentences, and in those sentences counts the words of 3 or more syllables.

4.1.3 Advanced Textual Feature Extraction

Once grammatical based features were extracted, advanced text processing was then applied to get actionable insights from the text.

A dictionary-based approach was performed creating a learning-corpus based on the content from the master programs on Dalarna University webpage. The dictionary-based algorithm uses a simple approach to segment text by scanning the words from the motivation letter with the words from the dictionary. If the series of words are found in the dictionary, the words are matched.

The common feature extraction procedure for text, known as Bag of Word (BOW) approach, was applied. The BOW in figure 8 is a representation of text that describes the occurrence of words within a document.

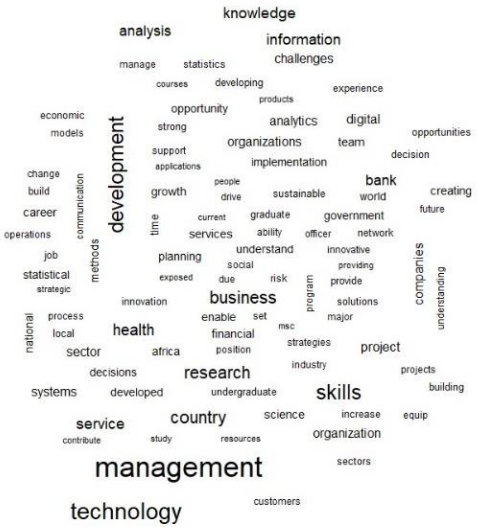


Figure 5 Representation of Bag of Words through a word cloud.

Another approach taken was with the Term Frequency-Inverse Document Frequency (TF-IDF), which is “a statistical measure that evaluates how relevant a word is to a document in a collection of documents” (Monkeylearn, 2020). The algorithm looks for how many times a word appears in a document and the inverse document frequency of the word across multiple documents.

The score returned by the TF-IDF algorithm, allowed the authors to count and plot unigrams and bigrams for each factor and its leveIs. The tokenization of consecutive sequence of words is called n-grams. In this case, it was decided to go for 1 and 2, hence the name unigram and bigram. It is then possible to know how a certain word is followed by another and build a model of relationships between them. The unigrams (Appendix 3) and bigrams (Appendix 4) were one of the methods that the authors used to try to identify the meaning of each factor.

4.2 Unsupervised Topic Modelling Methods

Several topics were expected to be covered by applicants in the motivation letters. These topics were unknown therefore the authors performed reverse engineering using unsupervised Topic Modelling methods.

‘In machine learning and natural language processing, a topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents’ (Wikipedia : The Free Encyclopedia, 2018). This method would allow the authors to identify the different topics covered in the motivation letters.

The methods explained below are considered as an unsupervised learning approach,

The first method considered was the Latent Semantic Analysis (LSA) which was the first topic mode patented in 1988. LSA is based on distributional hypothesis, which means that words which occurs in the same context have similar distribution thus similar meanings. Each motivation letter is treated as bag of words as the syntactic and semantic are ignored by this method. This method uses a single value decomposition on a document term matrix which means that the words frequency is being computed.

The workflow for LSA is largely the same for the Latent Dirichlet Allocation (LDA). “LDA is a mathematical method for estimating both of these at the same time: finding the mixture of words that is associated with each topic, while also determining the mixture of topics that describes each document” (Silge & Robinson, 2016). This method also

treats documents as bag of words with the difference that words are being assigned a probability of belonging to a topic. Since LSA does not consider any topic distribution, the LDA method was selected, as results are better defined (Pascual, 2019).

4.2.1 Cross Validation of Topic Modelling

Cross Validation of Topic Modelling helps identifying the optimal number of topics in a corpus of documents. R provides a library and package called "ldatuning" which computes 3 metrics to select the most optimal number of topics. Multiple LDA models are required to be trained to select the one with the best performance. The package was set with the "Gibbs" algorithm which provides the frequency of a topic in a document and the frequency of words in a topic. Figure 9 shows the most optimal number of topics from Griffith2004 metric which is evaluated to 14 topics whereas Arun2010 metric evaluated the optimal number of topics to 18 and CaoJuan2009 to 25.

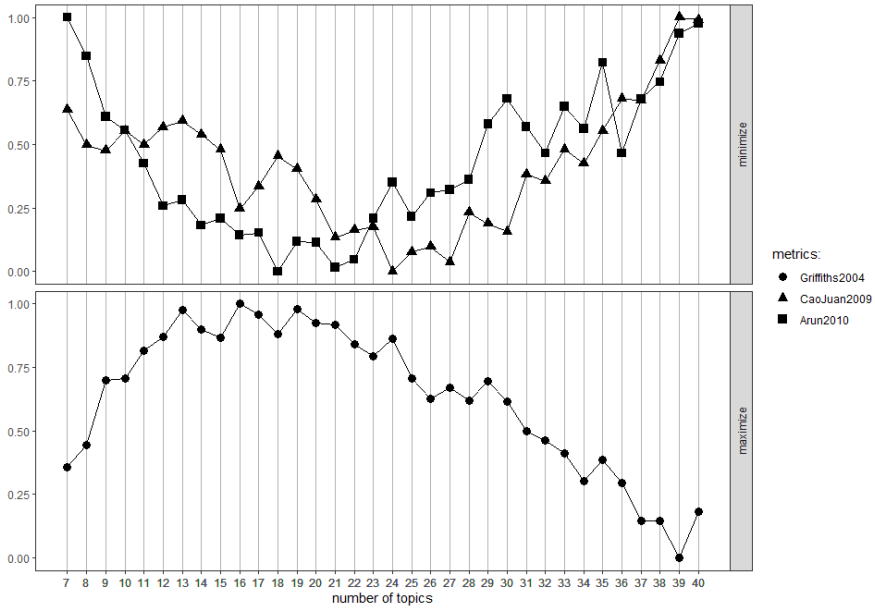


Figure 6 Tuning for Optimal Number of Topics

The reliability of the model was checked using Cross Validation and result on the optimal number of topics appears to be somehow identical.

Below figures display a topic similarity plot (Figure 10) across topics and the topic distribution for each motivation letter was identified (Figure 11). Although, the optimal number of topics identified with cross-validation is 14, the number of topics was reduced

to 10 as some topics were overlapping. It was expected that applicants were not constrained to 7 topics as they could freely cover any extra topics. The authors wanted to capture these extra topics although no factors related to these topics were graded by the program managers. The goal was to gain better understanding and insight on what is being covered in the motivation letters.

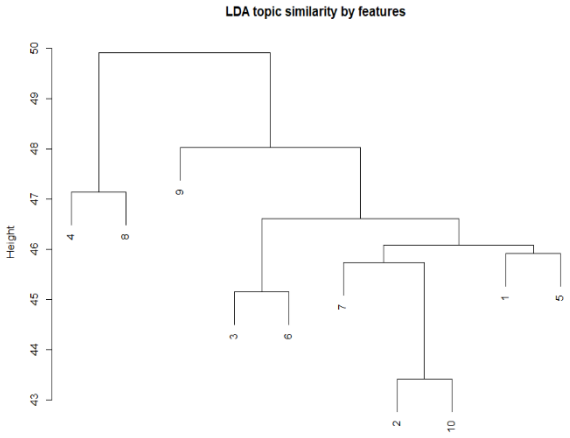


Figure 7 LDA topic similarity

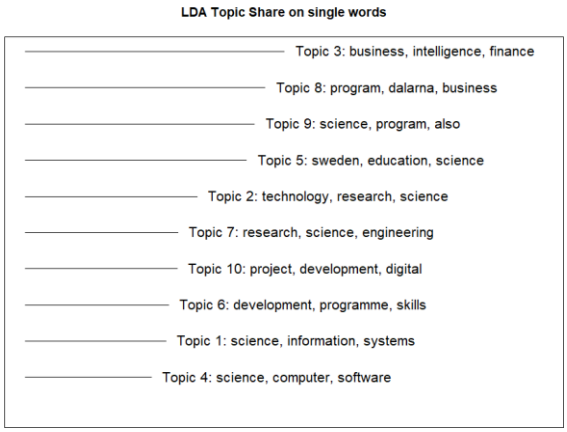


Figure 8 LDA Topic Share on Single Words

4.3 Supervised Learning Approach

Another approach that the authors discussed, involved doing reverse engineer of the factors through a supervised learning approach. "Supervised Learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs" (Russel & Norvig, 2009). If a good accuracy would have been achieved from the prediction models, it would have been possible, in theory, to identify what variables lead to that result and try then to understand how they were connected to the factors. Hence, it would have allowed the authors to specify the factors.

Before applying the supervised learning approach, a correlation analysis was done to quantify the association between the extracted features and the different factors graded by the program managers. Multiple variables were highly correlated with each other, and they were removed. This is done because correlated features in general don't improve models and they affect specific models in different ways. For example, multicollinearity

could yield solutions that vary and are numerically unstable. For Random Forest, highly correlated features can mask interactions between variables.

The next step was to identify the most important features, and that was done by using an Ordinary Logistic Regression (Appendix 4) and a Random Forest (Appendix 5). The features selected for each factor, were mainly retrieved from both the Ordinary Logistic Regression, the Random Forest algorithm and from the authors own perspective and understanding of the problem. This process was repeated for each Factor, which led to the creation of a model per factor.

The ordinary logistic regression is a powerful classification method, that like linear regression, relies on a model which relates the predictors with the outcome. In this case, 7 ordinary logistic regressions were applied, one for each factor. The Random Forest algorithm is no more than a multitree approach, where trees are based on separating data into subsections, through splits of the predictors. The Random Tree combines the result of the multiple trees to improve performance.

Once the model with the impactful variables were established the data set was divided between 75% as a training set and a 25% testing set. The different levels of each factor were set as the target variable and the algorithms Naïve Bayes Algorithm (NBA), Random Forest (RF) and Support Vector Machine (SVM) were used, first for classification and then for prediction purposes.

The Naïve Bayes Algorithm is a popular method for text categorization and derives from the Bayesian classifier, which has some simple principles. For a record to be classified, it is necessary to find all the data points with the same predictor profile, then, to determine what classes the records fit, and which class is the most predominant. For the Naïve Bayes Classifier, it is assumed that the value of the features is independent from any other feature, given the class variable. It is due to this strong assumption, that the algorithm is known as Naïve Bayes Classifier.

The Support Vector Machine algorithm (SVM) is a supervised method that can be used for classification and regression analysis. For classification purposes, an SVM algorithm is tasked to determine which category a new data point belongs by separating the data into two categories and identified by a gap that is as wide as possible. The new data point would then fall in one side of the gap and be attributed to that class.

5. Results

The result section covers the results from the data exploration and the results from the data analysis on the supervised learning approach. Data exploration is primordial in any kind of project to gain insights and to have a better understanding of the problem being analysed. As it was mentioned in earlier section human graders are subject to several outside forces which can impact their assessment. Therefore, the authors wanted to evaluate and understand the consistency of the factors being scored.

5.1. Data Exploration

An interactive approach to data exploration was performed by deploying a Shiny application. Shiny is an R package for building interactive outputs with friendly user interface which mainly requires a user to select from the drop down the Y and X metrics and attributes to be plotted.

The interactive interface provided several understandings from the dataset. The figure 5 shows that the sample data is relatively small with 107 motivation letters for BI program and 117 motivation letters for DS program thus the dataset was not segregated by program application. Despite DS program having 10 more applications it is noticeable that there is a considerable gap on the lowest score. It seems that DS program attracted a lower quality of motivation letters.

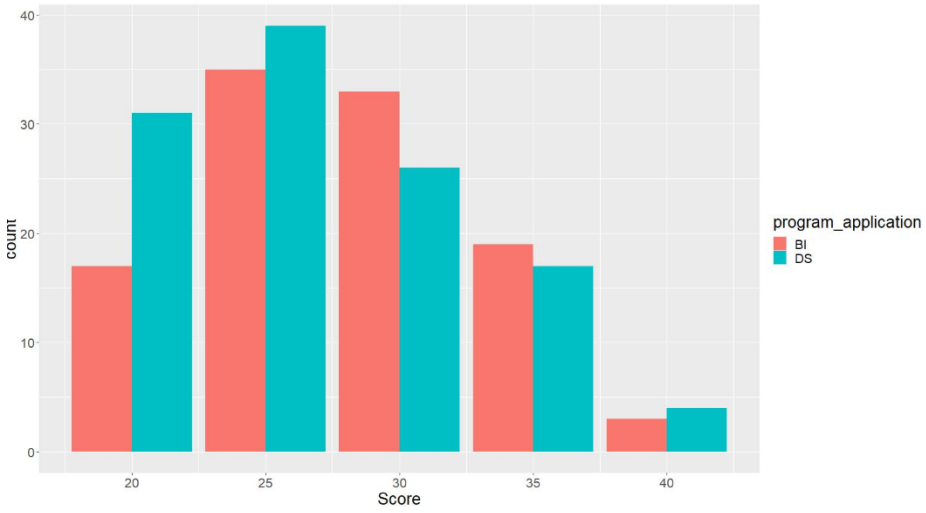


Figure 9 Distribution of score by program

Combining BI and DS applications, 21.72% of motivation letters received a score of 20, 33.48% of motivation letters received a score of 25, 26.24% of motivation letters received a score of 30, 15.84% of motivation letters received a score of 35 and only 2.71% of motivation letters scored 40.

Among the applicants 27% were females, 72% were males and the remaining candidates were not identified due to unavailable identification number. The figure 6 represents the distribution of scores by gender. The ratio of the genders in respect to their population is fairly distributed thus, we can assume that gender did not influenced the program managers in the scoring (Appendix 2).

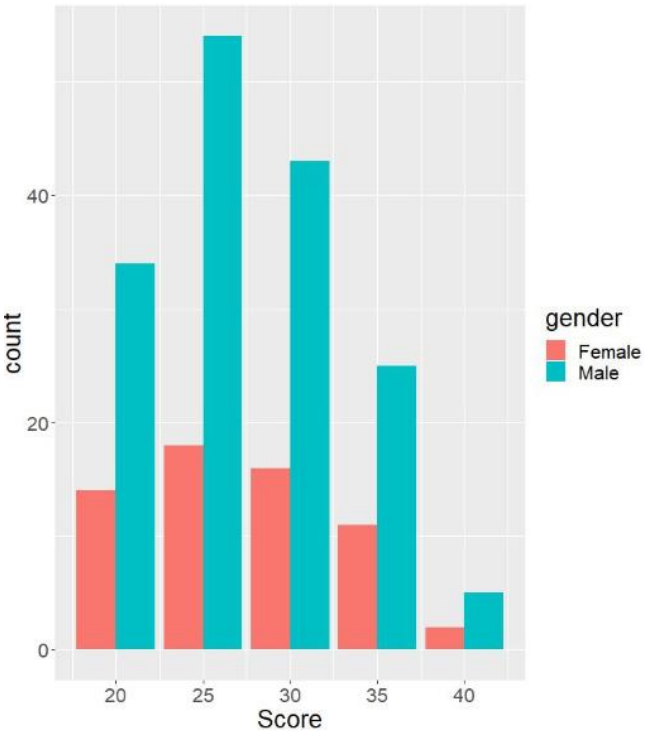


Figure 10 Distribution of applicant's gender by score

As mentioned in the data description, it was assumed that the factors were independent among them. To test our hypothesis, a correlation analysis took place and the result is displayed in figure 7. The method used was the same as for the result of figure 1. The data was condensed, the different levels were considered numbers, and then it was applied a Pearson correlation coefficient, which is a statistic method that measures the linear correlation between two variables. When two variables are independent, their correlation is zero, but if a calculated correlation is zero, it does not necessarily mean that they are independent.

Figure 11 Correlation plot between Factors and Score (here as merit rating)

Apart from the correlation obtained between Factor 6 and Factor 1, all the values were statistically significant. It was expected that the score (merit rating in the figure) would be correlated to the different factors, and it confirms our suspicions that different factors have a different weight on the final score. From figure 7, it is possible to select Factor 7 as the most correlated with the score, followed by Factor 3 and Factor 4. Also, to note that there seems to exist a mild correlation between Factor 7 and Factor 3. Lastly, the Factors are not independent, although most of them have statistically low correlations, there are a few which show a relative mildly strength. As text from a motivation letter follows a certain structure where some point might be interconnected or even overlap in some parts.

5.2. Data Analysis

Once the models were set in place (Appendix 8), the accuracy of the Support Vector Machine, Naïve Bayes and Random Forest were identified. The data was divided into a training and testing setting, with a partition of 75% and 25% respectively. The 7 different factors were used as the target variable, one after the other. A confusion matrix helped to understand the performance of the algorithms by reporting the false positives, false negatives, true positives, and true negatives of the predictions returned by the different algorithms.

Moreover, the Cohen's weighted kappa value was then computed to compare the overlap or the degree of agreement between the human ranking and the computer ranking.

Weighted kappa value is computed as follow:

$$k = 1 - \frac{\sum_{ij} wijOij}{\sum_{ij} wijEij}$$

Equation 1 Cohen's Weighted Kappa Equation

Where O_{ij} are the observed probabilities, $E_{ij} = p_i q_j$ are the expected probabilities and w_{ij} are the weights (with $w_{ji} = w_{ij}$).

Kappa result (Appendix 5) can vary between -1 and 1, where values below 0 indicates no agreement whereas values closer to 1 indicates the existence of an agreement.

Applying the algorithms on the grammar-based features, SVM, NBC and RF shows a very high accuracy of 96% on Factor 1. SVM showed an accuracy of prediction between

	SVM		**NBC**	
Factors	Accuracy	Kappa	Accuracy	Kappa
Factor 1	0.91	-0.01	0.9134	0
Factor 2	0.66	0.24	0.6	0.043
Factor 3	0.45	0.14	0.44	0.11
Factor 4	0.47	0.14	0.45	0.08
Factor 5	0.52	-0.05	0.57	-0.01
Factor 6	0.46	0.08	0.47	0.03
Factor 7	0.5	0.21	0.47	0.05

Table 1 Accuracy and Kappa for predictions done with Grammatical based features

50% and 60% on Factor 2, Factor 5 and 6 whereas Random Forest shows greater accuracy on Factor 2, Factor 3, Factor 4 and Factor 7 also between 50% and 60%. Although some Factor predictions shows high or average accuracy, the Kappa score appears to be extremely weak as it is always below 0.4 which means that only 15% to 35% of the data

are reliable. Since the numbers of observations in different classes vary greatly, the confusion matrix yields misleading results due unbalanced dataset.

RF was not applied on the Advanced Feature due to technical issues, nevertheless SVM and NBC were applied and did not demonstrate any performance enhancements. This went against the expectations of the authors, who were hoping to see some differences between the two.

	SVM		**NBC**		**RF**	
Factors	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
Factor 1	0.96	0	0.96	0	0.96	0
Factor 2	0.62	-0.0519	0.6	0.027	0.62	0.16
Factor 3	0.54	0.25	0.52	0.27	0.61	0.38
Factor 4	0.48	0.16	0.41	0.125	0.52	0.26
Factor 5	0.64	0.147	0.7	0.3205	0.625	0.24
Factor 6	0.5	0.0032	0.48	0.07	0.48	0.1
Factor 7	0.41	-0.0221	0.46	0.2	0.52	0.21

Table 2 Accuracy and Kappa for predictions done with Advanced Features

These results did not help the authors identifying what is behind each Factor. Nevertheless, due to the time they spent exploring the data and different algorithms, the authors came to suspect that certain factors may represent certain topic. Naturally, as they are only suspicious, they carry no scientific value and are more based on their own feeling.

Therefore, the authors suspect that Factor 1 relates to specificity of the letter. Is the letter generalist or targeted? Is the letter addressed at Dalarna University? Is it addressed at one of the masters? It is believed it is, especially because in this Factor only values of 1 and 2 were present, and none is a 0. Second, Factor 4 and 5, are suspected to relate to the academic and professional story of the applicant. Factor 7 is a statement where the applicant should explain how the programme will benefit his/her future career. All these suspicious are based on what is available in the website of Dalarna University for both programs. Although authors are aware that these points are most certainly present in the letters, these were not identified. Hence, future work is needed.

6. Conclusion

The algorithms results did not yield consistent accuracy and good Kappa scores and therefore failed to predict what is behind each Factor. Nevertheless, the authors still suspect of some factors based on the time that they invested analysing the data and working the algorithms. Also, the Grammatical based-features and the Advanced textual features which were cited in the literature review did not enhance the expected result in contrary to the authors expectation. The authors would have liked to run a deep-learning model but given that would be important to understand the factors, running a black-box was not seeing as a very interesting addition to the project.

The goal of this project would be to assist the program managers, by helping to create a program capable of emulating the human scoring of motivation letters. Although it has failed on its primary goal, the code pipeline was created, and it is available on request. The authors created the necessary code to extract text from images, store it in text documents and proceed with the exploratory data analysis.

The authors identified and discussed the limitations of the project, which had an impact on the overall project. The first limitation was related to image processing with Tesseract. The image processing was not done without some errors which resulted in loss of valuable information. The authors found it difficult to capture the core of the cover letter and tuning the parameters proved to be a complex and time-consuming activity. Since motivation letters are initially submitted by the candidate as text format, the retrieval of these files from a source system could ideally keep the identical format instead of converting these files into images.

The second limitation off this research is the size of the dataset. Only 221 motivation letters were provided while the accuracy could be improved with a larger dataset. From professional experience, the authors believe that the number of motivation letters should not be less than 500.

Third, since the factors were graded by 2 different program managers it is suspected that the Factors might suffer from to subjectivity. Human behaviour is affected by its environment thus it is not surprising that human graders are said to be inconsistent and unreliable (Zupanc, 2018). It is believed that biased scoring happens due to several aspects of the reader characteristics, like rating experience, reader psychology (internal

factors that affect the reader) and rating environment like pressure (Brent, 2013). Therefore, human nature introduces variance into the scores and impacts their validity (Susan M. Lottridge, 2013).

The authors did not manage to capture the meaning of the factors, but the path is open for further investigation. The topics were not identified accurately, but multiple hypothesis was found for the different topics. The authors believe that it is possible to extract more from these motivation letters, if for example, context can be explored. The authors did however manage to aggregate the data and organize it in a tidy way, which will facilitate further investigation for future researchers.

From the work developed along this thesis, many topics were identified that could lead to future work. Some of the topics the authors have discussed are:

1. Performance of the students on their academic life, and after graduation. Is there a relationship between the score each student obtained and their academic and professional path? Did students with a high grading cover letter, obtained a better academic performance? What students eventually dropped out of the course? The authors believe that these questions can provide an insight to the program managers and to the students themselves.
2. Is there a relationship between gender, type of writing, or any other hidden correlation? For example, culture and drop-out ratio. Do students from a specific country seem to struggle in graduating?
3. What results ad insights would be possible to obtain by using deep learning?
4. How can we make use of context to improve topic identification?

Appendix 1: A comparison of the key features of the state-of-the-art AEE systems

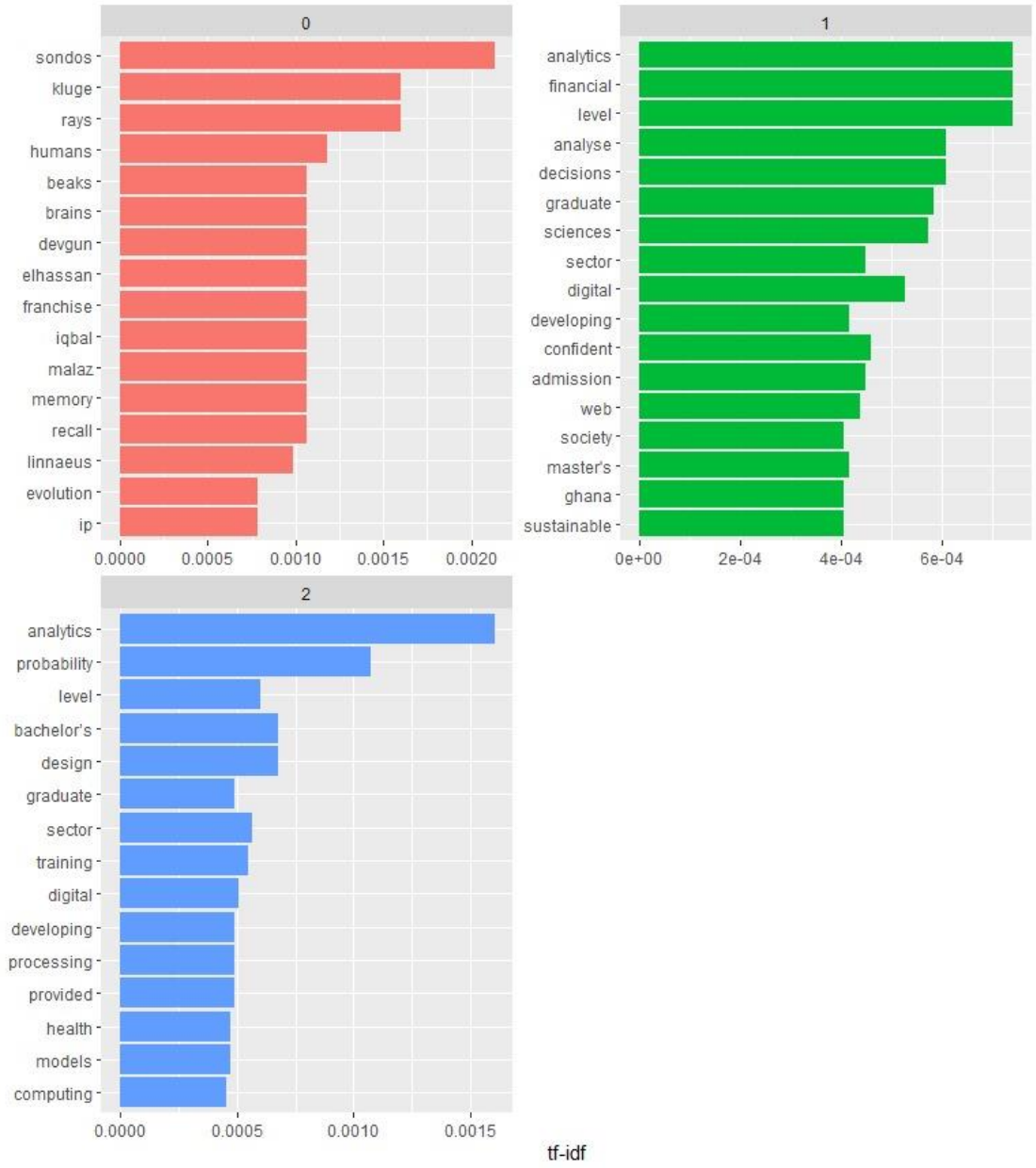
A comparison of the key features of the state-of-the art AEE systems.

AEE System	Attr. Types	Methodology	Prediction Model
PEG [Page]	Style	Statistical	multiple linear regression
PS-ME [Mason and Grove-Stephenson]		NLP	linear regression
e-rater [Burstein et al.]	Style & Content	NLP	linear regression
IntelliMetric [Schultz]			multiple mathematical models
Bookette [Rich et al.]			neural networks
OzEgrader [Fazal et al.]			
CRASE [Lottridge et al.]			machine learning
AutoScore [Shermis and Hamner]			statistical model
Lexile [Smith et al.]			Lexile measure
Ranked-based AEE [Chen et al.]			learning to rank
Multi-classifier Fusion AEE [Bin and Jian-Min]			ensemble classifiers
BETSY [Rudner and Liang]		Statistical	Bayesian networks
SEAR [Christie]			linear regression
Neural Essay Assessor [Taghipour and Ng]		Deep learning	recurrent neural networks
AES using NN [Alikaniotis et al.]			
AEG using MN [Zhao et al.]			memory networks
LightSIDE [Mayfield and Rosé]	Content	Statistical	machine learning
IEA [Foltz et al.]		LSA, NLP	
Semantic-tree-based AEE [Chali and Hasan]		LSA, tree kernel functions	cosine similarity
GLSA based AEE [Islam and Hoque]		GLSA	
Markit [Williams and Dreher]		NLP, PMT	linear regression
SAGrader [Brent et al.]	Semantic	FL, SN	rule-based expert systems
OBIE-based AEE [Gutierrez et al.]		OIE, DL	/
SAGE		OIE, NLP	random forest

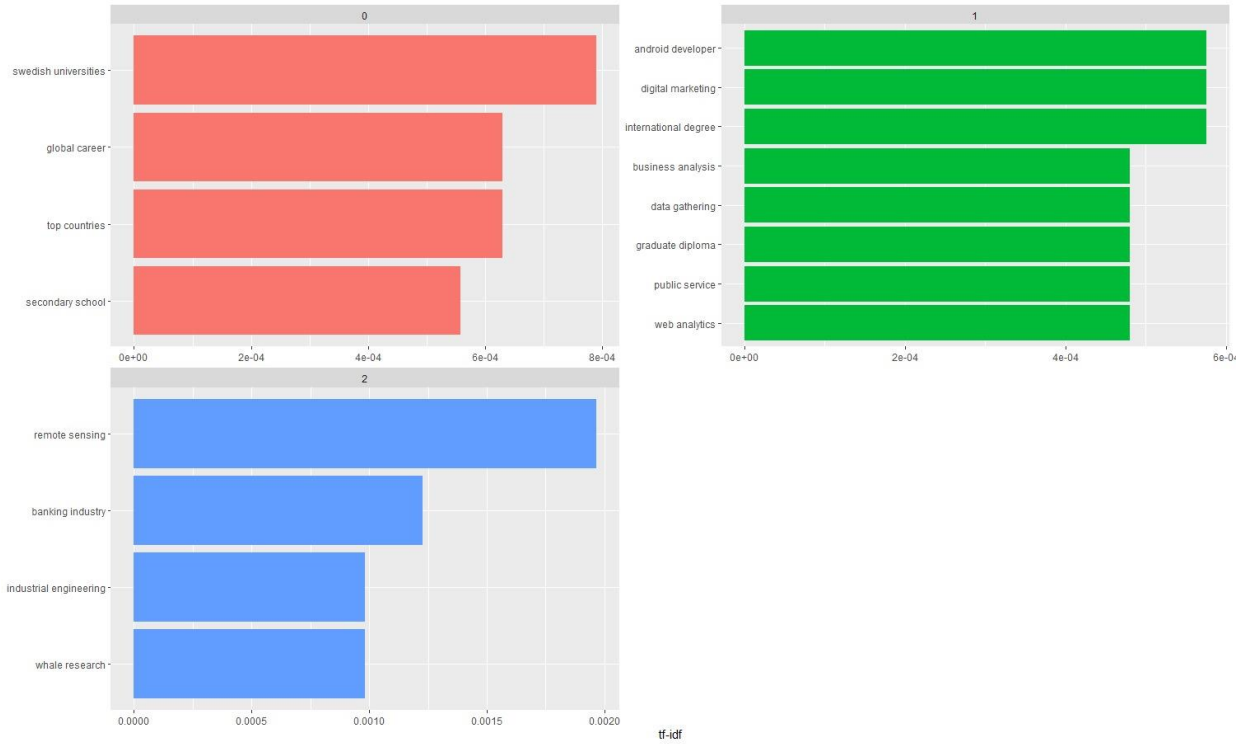
Appendix 2: Ratio between Male and Female by Score

Score	*Female*	*Male*
20	23%	21%
25	30%	34%
30	26%	27%
35	18%	16%
40	3%	3%

Appendix 3: Example of Unigrams



Appendix 3: Example of Bigrams



Appendix 4: Ordinal Logistic Regression Output

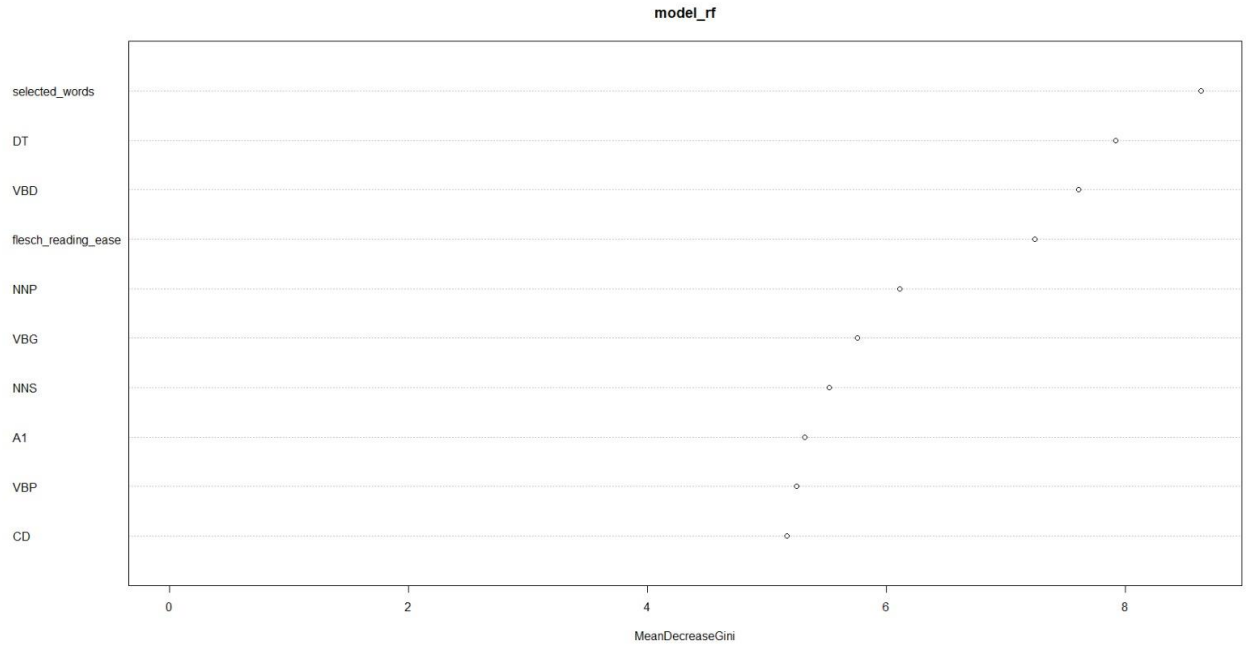
```
Call:
polr(formula = Factor5 ~ . - ID_image, data = new_df_factor,
    Hess = TRUE, method = "logistic")

Coefficients:
                              Value Std. Error  t value
prio                       -0.456172   0.136304 -3.34672
flesch_reading_ease        -0.023174   0.018330 -1.26424
lexical_diversity_root_TTR -0.061308   0.240212 -0.25523
count_misspelled            0.005973   0.032128  0.18591
RP                         -0.114305   0.167470 -0.68254
VBD                         0.011887   0.029129  0.40809
PDT                         0.144257   0.259236  0.55647
VBN                        -0.033903   0.040129 -0.84483
WDT                        -0.054382   0.089700 -0.60626
NNPS                        0.137177   0.144755  0.94765
PRP_1                       0.002920   0.029001  0.10068
JJS                         0.050413   0.112931  0.44640
DT                          0.009808   0.017147  0.57200
A1                          0.003002   0.020372  0.14737
JJR                         0.365828   0.112386  3.25511
NNP                         0.003804   0.008037  0.47328
VBZ                        -0.095478   0.043812 -2.17926
MD                         -0.119836   0.053656 -2.23339
A2                         -0.334358   0.116373 -2.87316
VBP                        -0.067343   0.041693 -1.61523
WP                         -0.004777   0.166312 -0.02872
CD                          0.057007   0.046299  1.23129
RB                          0.010854   0.029130  0.37261
WRB                         0.053093   0.082034  0.64721
NNS                         0.040150   0.018923  2.12179
VBG                         0.053192   0.032766  1.62336
selected_words             -0.006734   0.022669 -0.29705
extract1                   -0.336378   0.718542 -0.46814
extract2                    1.263266   1.008329  1.25283
extract3                   -0.054777   0.691805 -0.07918
extract4                    0.979436   0.932630  1.05019
extract5                    0.963300   0.781547  1.23256
extract6                    0.156144   0.857074  0.18218
extract7                    0.871687   0.783127  1.11308
extract8                    0.133460   0.889510  0.15004
extract9                    0.125496   0.730863  0.17171

Intercepts:
    Value   Std. Error t value
0|1 -3.1740  2.4036    -1.3205
1|2  0.5823  2.3953     0.2431

Residual Deviance: 334.1082
AIC: 410.1082
```

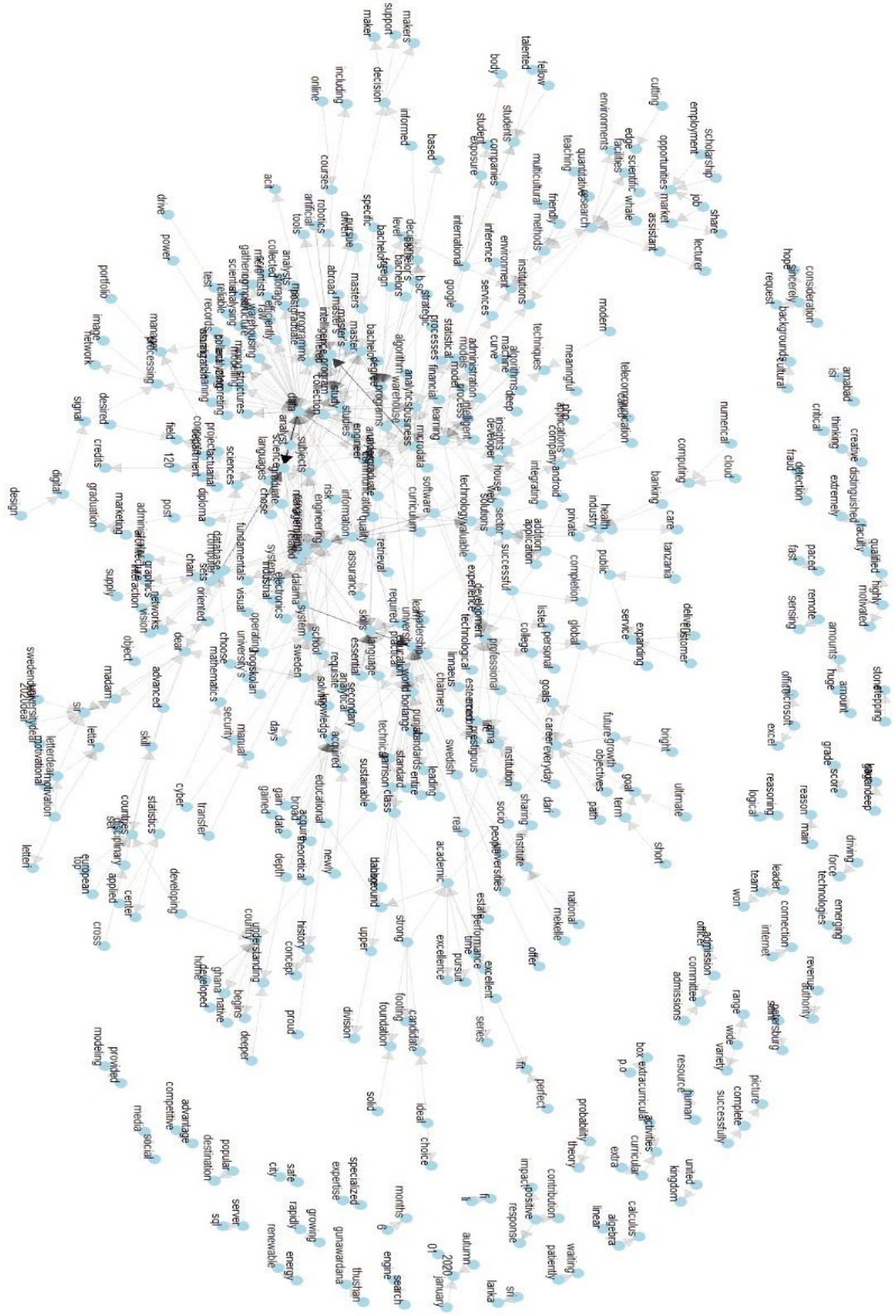
Appendix 5: Random Forest Output



Appendix 6: Kappa Results

Value of Kappa	Level of Agreement	% of Data that are Reliable
0-.20	None	0-4%
.21-.39	Minimal	4-15%
.40-.59	Weak	15-35%
.60-.79	Moderate	35-63%
.80-.90	Strong	64-81%
Above .90	Almost Perfect	82-100%

Appendix 7: Example of Word Network



Appendix 8: Models

Grammar Based Feature

Factor1: **df_model <- df_factor %>% select(Factor1 , selected_words, prio, lexical_diversity_root_TTR)**

Factor2: **df_model <- df_factor %>% select(Factor2 , flesch_reading_ease, NNP, VBD, count_misspelled)**

Factor3: **df_model <- df_factor %>% select(Factor3 , prio, flesch_reading_ease, WRB, MD, lexical_diversity_root_TTR, NNP, selected_words, count_misspelled)**

Factor4: ***df_model <- df_factor %>% select(Factor4, prio,flesch_reading_ease, JJR, MD, lexical_diversity_root_TTR, NNP,sentence_count)***

Factor5**: df_model <- df_factor %>% select(Factor5 , selected_words, prio, flesch_reading_ease, DT, NNP)**

Factor6: **df_model <- df_factor %>% select(Factor6 , VBD, RB, flesch_reading_ease, lexical_diversity_root_TTR, NNPS)**

Factor7**: df_model <- df_factor %>% select(Factor7 , prio, PRP_1, flesch_reading_ease, lexical_diversity_root_TTR, selected_words)**

7. References

Akrami, N., Fernquist, J., Isbister, T., & Lisa Kaati, P. B. (2019). *Automatic Extraction of Personality from Text.* Uppsala: Uppsala University, Swedish Defence Research Agency.

Alikaniotis, & all., e. (2016). Automatic Text Scoring Using Neural Networks. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistic* (pp. 715-725). Berin, Germany: Association Computational Linguistic.

Arthur, G., Danielle, M. N., Max, L., & Zhiqiang, C. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*(2), 193–202.

Boiangiu, C.-A. &.-C. (2016). Voting-Based OCR System. *Journal of Information Systems & Operations Management*, 10.

Brent, B. (2013). *Handbook of Automated Essay Evaluation: Current Applications and New Directions.* (M. D. Burstein, Ed.) New York: Routledge,.

Burstein, M. D. (2003). Automated essay scoring: A cross-disciplinary perspective. *Lawrence Erlbaum Associates, Manwah*, xiii–xvi.

D. Shermis, M., Burstein, J., & Bursky, S. A. (2013). *Handbook of Automated Essay Evaluations: Current Applications and New Directions.* New York: Routledge.

Dalarna University. (2020). *Master in Data Science - Dalarna University*. Retrieved May 16, 2020, from Dalarna University: https://www.du.se/en/study-at-du/programmes-courses-and-course-packages/programmes/data-science-masters-programme/

H. Schwartz, J. E. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *8*(9), p. e73791.

Harshanthi, G. (2019). *Automated Essay Evaluation using Natural Language Processing and Machine Learning.* Columbus: Columbus State University.

Hoffstaetter, S. (2020, 16 02). *Pytesseract*. Retrieved from https://pypi.org/project/pytesseract/

Holtzscher, M. (2019, January 29). *spacy-readability 1.4.1*. Retrieved from Pypi: https://pypi.org/project/spacy-readability/

Jr, R. R. (1999). A fiver-factor theory of personality. In *Hadnbook of personality: Theory and research, vol. 2* (pp. 139-153).

Kagan, J. (2007). "A trio of concers",. In *Perspectives on physchological science, vol 2* (pp. 361-376).

Ke, Z., & Ng, V. (2019). Automated Essay Scoring: A Survey of the State of the Art. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, (pp. 6300-6308).

Lopez, M., & Kalita, J. (2017). Deep Learning applied to NLP. .

Madnani, N., & Cahill, A. (2018). Automated Scoring: Beyond Natural Language Processing. *Proceedings of the 27th International Conference on Computational Linguistic* (pp. 1099-1109). Santa Fe, New Mexico, USA : Association for Computational Linguistic.

Monkeylearn. (2020, 05). Retrieved from https://monkeylearn.com/blog/what-is-tf-idf/

Östling, R., Smolentzov, A., Tyrefors Hinnerich, B., & Erik, H. (2013). Automated Essay Scoring for Swedish. *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 42-47). Atlanta, Georgia: Association for Computational Linguistics.

Page, E. B. (1966). The Imminence of Grading Essays by Computer. *Phi Delta Kappan, 47*(5), 238-243.

Pascual, F. (2019, September 26). *MonkeyLearn*. Retrieved from Introduction to Topic Modeling: https://monkeylearn.com/blog/introduction-to-topic-modeling/#examples-of-topic-modeling-and-topic-classification

Pennebaker, J., & King, L. (1999, 12). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*(6), 1296-1312.

Russel, S. J., & Norvig, P. (2009). *Artificial Intelligence: A Modern Approach* (3 ed.). (P. Education, Ed.) Boston.

Silge, J., & Robinson, D. (2016). *Text Mining with R*. Retrieved from Tidy Text Mining: https://www.tidytextmining.com/topicmodeling.html

Singh, S. (2019, July). *How to Get Started with NLP – 6 Unique Methods to Perform Tokenization*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2019/07/how-get-started-nlp-6-unique-ways-perform-tokenization/

Smolentzov, A. (2013). *Automated Essay Scoring.* Faculty of Humanities, Department of Linguistic. Stockholm: Stockholm University.

Study at Dalarna Univeristy. (2020). *Business Intelligence: Master Programme*. Retrieved May 16, 2020, from https://www.du.se/en/study-at-du/programmes-courses-and-course-packages/programmes/business-intelligence-master-programme/

Susan M. Lottridge, E. M. (2013). *Handbook of Automated Essay Evaluation: Current Applications and New Directions.* (M. D. Burstein, Ed.) New York: Routledge.

Wikipedia : The Free Encyclopedia. (2018, June). *Machine learning*. Retrieved from Wikipedia : The Free Encyclopedia: https://en.wikipedia.org/wiki/Machine_learning

Woods, B., Adamson, D., Miel, S., & Mayfield, E. (2017). Formative Essay Feedback Using Predictive Scoring Models. *the 23rd ACM SIGKDD International Conference*, (pp. 2071-2080).

Young, T., Hazarika, D., Poria, S., & Erik, C. (2018). Recent Trends in Deep Learning Based Natural Language Processing. 55.

Zupanc, K. (2018). *Dissertation: Semantics-based automated essay evaluation.* Ljubljana: University of Ljubljana: Faculty of Computer and Information Science.