Towards Next Generation Rubrics: An Automated Assignment Feedback System

Nilupulee Nathawitharana
La Trobe University
nathawitharana.n@students.latrobe.edu.au

Qing Huang
La Trobe University

Kok-Leong Ong
La Trobe University

Peter Vitartas
La Trobe University

Madhura Jayaratne
La Trobe University

Damminda Alahakoon
La Trobe University

Sarah Midford
La Trobe University

Aleks Michalewicz
La Trobe University

Gillian Sullivan Mort
La Trobe University

Tanvir Ahmed
La Trobe University

Abstract

As the use of blended learning environments and digital technologies become integrated into the higher education sector, rich technologies such as analytics have shown promise in facilitating teaching and learning. One popular application of analytics is Automated Writing Evaluation (AWE) systems. Such systems can be used in a formative way; for example, by providing students with feedback on digitally submitted assignments. This paper presents work on the development of an AWE software tool for an Australian university using advanced text analytics techniques. The tool was designed to provide students with timely feedback on their initial assignment drafts, for revision and further improvement. Moreover, it could also assist academics in better understanding students' assignment performance so as to inform future teaching activities. The paper provides details on the methodology used for development of the software, and presents the results obtained from the analysis of text-based assignments submitted in two subjects. The results are discussed, highlighting how the tool can provide practical value, followed by insights into existing challenges and possible future directions.

Keywords: Learning Analytics, Automated Writing Evaluation, Text analysis, Assignment feedback

1 Introduction

With the emergence of big data and advances in text analytics, information retrieval, machine learning and natural language processing, academics are able to explore new opportunities to understand student performance in assignment submissions. This is of particular importance to academics involved in extra-large subjects and Massive Open Online Courses (MOOCs).

While much of this work is still being advanced in a range of disciplines, the opportunity arises to explore how such techniques can assist and be applied to the management of assignment marking and moderation in large subjects. The adoption of digital submissions for assignments via the Learning Management System (LMS) and the development of text-analytic algorithms has seen a range of Automated Writing Evaluation (AWE) or Automated Essay Scoring (AES) systems emerge (Deane, 2013; Shermis & Burstein, 2003), which have been shown to be comparable to human markers (McNamara, Crossley, Roscoe, Allen, & Dai, 2015).

Examples of automated analysis and evaluation of written text are emerging in a variety of contexts, from formative feedback in writing instruction (from primary through tertiary education), to summative assessment (e.g. grading essays or short answer responses with or without a second human grader). As class sizes increase, there is a corresponding increase in the use of large-scale exams and the generation of large volumes of writing to be evaluated and assessed (e.g. 1000+ students in first year university classes, NAPLAN in Australia, exams based on the Common Core State Standards Initiative in the US and the rise in popularity of MOOCs). At the same time, the increase in interest of marketers to explore the meaning of messages and brand sentiment in social media such as Facebook, Twitter and blogs has seen increased demand for ever-more sophisticated text analysis tools. Such developments in text analysis are finding their way into applications for education and demonstrate potential for the assessment of text based writing such as assignments.

As a result, AWE systems have drawn on multidisciplinary insights from computer science, linguistics, writing research, cognitive psychology, and Educational Data Mining (EDM). In this research, the use of business analytics methods is being applied to the development and refinement of software that can not only provide formative feedback to students, but also assist teachers to understand their students' performances and areas where guided instruction could be directed to understanding terms and concepts covered, or not covered, in assignments.

1.1 Next Generation Rubrics

The Next Generation Rubrics (NGR) project (Vitartas et al., 2016) was established at an Australian university as a collaboration between a small number of academics and a newly appointed business analytics team. The project was supported by an internal Learning and Teaching grant and sought to develop a tool that provided students and academics with information about the performance of text based assignments as a proof of concept.

The NGR software aims to provide descriptive timely feedback to students on their digitally submitted text assignments as well as assist the academic to better understand students' assignment performance and learning outcomes. The starting point for the development of the NGR software was a marking rubric, as these underpin many standardised marking schemes at universities. While it is acknowledged that rubrics have come to have a range of meanings to various people (Dawson, 2017), in this paper its meaning is based on Popham's definition: "a scoring guide used to evaluate the quality of students' constructed responses" (Popham, 1997, p.72) and consists of an evaluative criterion, and guidance on expectations for associated scores or marks (Popham, 1997).

The project analysed assignments from two subjects. The first is an introductory first-year subject in the Bachelor of Arts (BA). The assessment regime required students to submit a 1500-word assignment and a total of 115 assignment scripts were included for analysis. Students answered one of five questions, analysing the ways individualism, imperialism or secularisation has impacted contemporary society. The second subject is at the Masters level from the marketing discipline (MKT) in the Business School at the same university. Students were required to present a 2500-word report on a strategic analysis of a new product introduction for a video streaming company. A total of 80 assignment scripts were analysed from this subject.

The software was developed to evaluate students' performance for the purpose of providing feedback. Moreover, cluster analysis of the assignments was also conducted based on calculated evaluation measures in order to provide the academic with a detailed view of students' assignment performance, e.g. how evaluation measures are performed across the student cohort. By examining the results from the analysis, a greater understanding of the computer's ability to assess student performance could be attained, and the results could be compared to the human issued marks for each assignment.

The remainder of the paper first provides the background to the development of the AWE tool and its role in facilitating teaching and learning. The methodology used for NGR's development is then outlined, followed by the presentation and discussion of the results from this study. Furthermore, we also share the clustering results and discuss how they could assist academics to extract further insights into their cohort. Prior to the conclusion, we discuss the challenges and limitation in the field and suggest possible future directions.

2 Background

Innovative analytical tools are providing opportunities for educational designers and teachers to understand student performance in much greater detail than ever before. Tools such as text analytics, information retrieval, machine learning, natural language processing and learning analytics form part of the suite of big data analytics that has the potential to provide evaluative feedback on students' work (Shermis & Burstein, 2013).

In this section, we provide a brief review of some key concepts underlying the technology being applied in AWE, drawing on insights from computer science, linguistics, writing research, cognitive psychology, Educational Data Mining (EDM) and Learning Analytics (LA). We provide a synopsis of the AWE tools and then discuss how AWE tools could possibly assist both the students and the academic to gain valuable insights from text-based assignments.

2.1 Educational Data Mining (EDM) and Learning Analytics (LA)

The EDM community website describes EDM as "an emerging discipline, concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings, and using those methods to better understand students, and the settings which they learn in" ("Educational Data Mining," 2017). With the advent of MOOCs and publicly available data, such as the Pittsburgh Science of Learning Center DataShop, EDM research has accelerated in recent years.

Learning Analytics (LA) was defined in the first international Learning Analytics and Knowledge (LAK) conference as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs" (1st International Conference on Learning Analytics and Knowledge, 2010). It draws on the increasing range of data available from digital learning tools and environments.

Siemens and Baker (2012) promote a closer collaboration between the EDM and LA communities, as the two groups share the goals of improving both educational research and practice through the analysis of large-scale educational data. One interesting development from EDM/LA that is relevant to AWE is White and Larusson's 'point of originality' tool (White & Larusson, 2010). This is designed to help instructors in large university courses, such as first year gateway courses, to monitor students' understanding of key concepts. The system uses WordNet (Miller, 1995), a large lexical database of English words, to "track how a student's written language migrates from mere paraphrase to mastery, isolating the moment when the student's understanding of core concepts best demonstrates an ability to place that concept into his or her own words, a moment we've chosen to call the 'Point of Originality'" (White & Larusson, 2010, p. 817).

2.2 Natural Language Processing (NLP)

Put simply, NLP is "an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things" (Chowdhury, 2003, p. 51). Some of these 'useful things' include machine translation, speech recognition, information retrieval and extraction, summarisation, and, relevant to AWE, text processing. Liddy (2001) points out that NLP can operate at various levels of linguistic analysis, including phonology, morphology, lexical, syntactic, semantic, discourse and pragmatic.

2.3 Automated Writing Evaluation (AWE)

AWE systems can be classified as either simulation-based assessments or response-type systems (Williamson, Xi, & Breyer, 2012). The former present computerised simulations of real-life scenarios, and are usually specific to a certain test (such as the United States Medical Licensing examination). The latter are more generalisable in that they score a typical type of response such as mathematical equations, short written responses, spoken responses, or essays. Essay scoring has been a particular focus for many automated systems and numerous essay evaluation systems are now used in formative feedback as well as high-stakes testing. In these tests the automated assessor acts either as a second rater to assist human scorers (e.g. in ETS' TOEFL test - ETS 2015) (ETS®, 2017b), or as the sole rater (e.g. the Pearson Test of academic English (PTE Academic) uses automated scoring for writing and speaking (Pearson, 2012)). The following section provides a synopsis of a number of available tools.

A history of AWE generally begins with the work of English teacher turned researcher Ellis Page, and his Project Essay Grade (PEG) beginning in the 1960s. In a 1966 article in *The Phi Delta Kappan*, Page insisted "we will soon be grading essays by computer, and this development will have an astonishing impact on the educational world" (Page, 1966, p. 238). Ellis had an optimistic vision for writing feedback being provided to students much more extensively and in a timely manner than could be achieved by English teachers in school or college.

Something of Page's vision is now closer to reality in various automated formative tools, offering immediate descriptive feedback on writing (e.g. WriteLab (WriteLab, n.d.), Turnitin's Revision Assistant (Turnitin, 2017), Pearson's WriteToLearn™ (Pearson, 2017), ETS' Criterion® (ETS®, 2017a), and Vantage Learning's MyAccess! (Vantage Learning, 2016)). In addition to these proprietary products, a number of freely available services from the academic domain are also available to examine text and extract phrases including Coh-Metrix (Coh-Metrix, n.d.), WordNet (Miller, 1995), TerMine (Frantzi, Ananiadou, & Mima, 2000), MALLET (McCallum, 2002), Stanford Core NLP (Manning et al., 2014) and Natural Language Toolkit (NLTK) (Bird, Klein, & Loper, 2009).

2.4 The Role of AWE Tools for Both Student and Teacher Insights

A considerable amount of work being undertaken in the development of AWE tools is focused on more accurately evaluating how students perform in their writing assignments. However, evaluation results are only as valid as when they are able to provide meaningful insights to both students and teachers in terms of informing future learning and teaching. Therefore, a critical role of AWE tools could be providing the opportunity to understand better how and why students are performing in this way.

In fact, the NGR project was initiated by the university's need to provide students with timely and personalised feedback coupled with the desire to extract further insight into students' assessment writing for academics. While most of the existing AWE tools focus on providing automated scoring rather than generating feedback, our project aims to cater to the specific requirements of the university and provide a customised solution to the need for high-quality feedback on written assessment as a formative learning tool.

Therefore, the research presented in this paper focuses on automatically providing descriptive feedback to students via the analysis of text-based assignments. Software was developed on the basis of five dimensions of analysis to provide formative feedback. Furthermore, clustering techniques were applied to uncover any hidden patterns in the text data that may contain useful information for enhancing an understanding of the results. A stronger understanding of students' assignment performance could potentially provide meaningful insights for academics, informing future teaching activities and subject design. The project is continuing and has already provided useful insights into students' writing performance and the effectiveness of teaching activities.

3 Methodology

3.1 NGR Software Functionality and Architecture

The NGR software assists both students and academic staff. The main functionality is assessing digitally submitted assignments and providing descriptive feedback to students. This process allows students to submit an unlimited number of draft versions of their assignments, obtain feedback and improve them before they make a final submission for (human) marking. Upon submission of an assignment, the software conducts content analysis, evaluates the assignment based on an evaluation criterion, and generates feedback based on the analytical results explaining which areas need further improvement. For academic staff, the software allows the performance of individual students and classes to be monitored. Academics can add subjects and assignments for their subjects using the software. In addition, they can set evaluation criteria for those assignments which will be used during the analysis process.

The software facilitates the analysis of different types of written assignments via a generalised evaluation mechanism. However, evaluation of assignments might depend on the nature of the writing task, and the importance of different evaluation measures might depend on the type of assignment and subject context (e.g. report or essay). Therefore, the software facilitates the functionality to customise evaluation measures based on the type of assignment. The software is accessed by both academic staff and students, and the user interface is implemented to be simple but user-friendly. More importantly, different functionalities need to be provided by the software for the different users. Students are able to access the analysis results and descriptive feedback generated for the assignments they have uploaded and the academic staff are able to view the performance of the class as well as query and observe the performance of individual students. The software caters for this requirement by maintaining different access levels for different types of users.

The high-level architecture of the NGR software is presented in Figure 1. The software consists of four layers; a user interface tier is the topmost level and defines the presentation to the users. This tier communicates with the business logic tier to retrieve essential information for the presentation. The business logic tier manages users, grading, subjects, assignments, and dashboard presentation. The various analytics are implemented on the service tier and the database tier stores assignments and other data essential for the analysis. During the analysis, various analytical functions implemented in the service tier communicate with the data tier to retrieve critical information, and afterwards the analysis output is presented to the business logic tier, which organises the results and presents them to the end-user via a user interface tier.

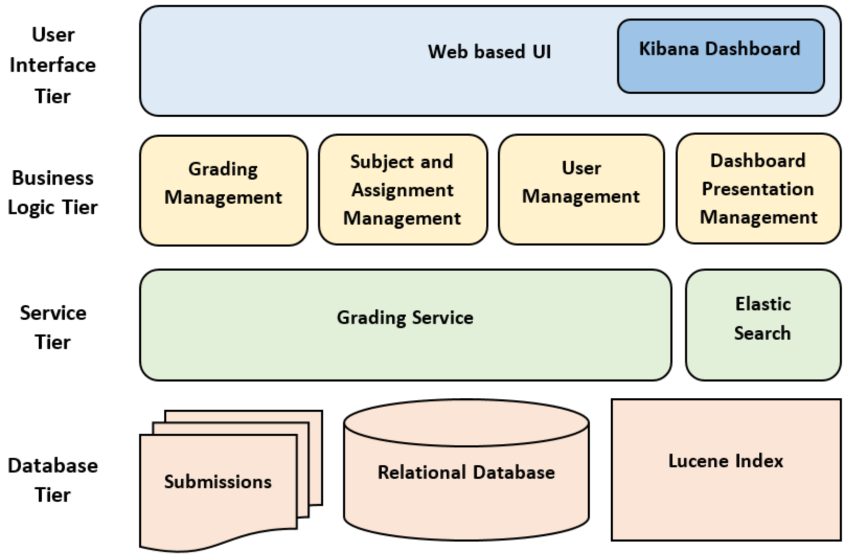


Figure 1: High-level architecture of the NGR software

The NGR project is ongoing and the software continues to be refined. The current version of the software facilitates the calculation and evaluation of measures and presents the outcome to dashboards. Academic staff can view the performance of classes and individual students. Likewise, students can access the software and view the values of the evaluation measures for the assignments they have uploaded. The weighting of evaluation measures and personalised feedback are currently under development and will be incorporated into the next version of the software.

The following subsections provide details of the implementation of the analytical component, core dimensions used for the evaluation, and tools and technologies utilised in the development of the software.

3.2 Implementation of the Analytical Component

Identifying proper evaluation criteria to be implemented for the assignment analysis logic has been one of the main objectives during the implementation phase of the NGR's analytical component. In addition, the evaluation criteria expected to assess assignments and identify the areas have been a focus of the software development, but require further improvement.

Rubrics which defined the scoring guide play a vital role for the marker during manual assignment grading. Therefore, the rubric was considered the starting point for the development of the analytical component. A set of already marked assignments was analysed to identify important evaluation measures, and for this project, assignments from two different subjects were used to identify valid evaluation criteria. The discipline and level of the degree were two controlling factors used for assignment assessment, and it was also noted that trialling subjects from different disciplines and degree levels provided insights into a wider range of issues which might occur during the analysis process. Therefore the two subjects selected for developing the analytical component belonged to different disciplines and degree levels.

The first subject was selected from the Bachelor of Arts (BA) degree and therefore represented the humanities disciplines. Several prominent concepts that emerged during the European Enlightenment (such as individualism, imperialism, secularisation and freedom/democracy) are explored from their origins to their manifestation in contemporary society in this first year core subject. The assignment selected from this subject was an essay of 1500 words based on five themed questions. Approximately 800 student essays were submitted to this subject, and one themed question with 115 submissions was selected for analysis. The second subject selected for the analysis was chosen from the Masters of Management Marketing (MKT). This subject aimed to provide guidance and develop the necessary skills to understand and implement marketing strategy in a business context. The selected assignment was a 2500 word executive report in which students conducted a strategic marketing analysis of a new video streaming company. There were 80 submissions of this assignment.

3.3 Core Dimensions for Assessment

As mentioned in the previous section, a marking rubric provides the expectations for each grade and describes different areas which need to be considered for the allocation of marks. In general, the marking criteria for an assignment includes multiple components, and the marking guide for the selected two assignments were consulted to identify important measures for the evaluation. It has been noted that the assignment structure, proofreading errors, evidence for research and critical thinking, use of discipline terms, and use of accurate referencing mechanism are important for mark allocations. Consequently, the following five dimensions were incorporated into the analysis.

1. Assignment statistics
2. Proofreading errors
3. Research, critical thinking, and discipline words and phrases
4. Assignment readability
5. Referencing information

Assignment statistics measure how well each assignment complies with the submission guidelines. The assignment guide included the submission requirements stating page and word limits, and final format for the submission. Word count, paragraph count, and page count are calculated as assignment statistics to verify whether an assignment follows the prescribed format. Proofreading errors, spelling and grammar error counts are calculated per 1000 words. Considering the length of the assignment provides a standardised representation of proofreading errors as the occurrence of those errors are proportional to the length of the document. Research, critical thinking and discipline terms are captured to verify whether the assignment covered required disciplinary concepts, contains research related content, and critically evaluated the evidence. The use of research, critical thinking, and discipline terms are calculated by referring to a set of term lists. Therefore during the initial development phase, subject experts were consulted to retrieve research and discipline terms for each assignment. For measuring critical thinking, a term list developed by Paul (1995) was used.

Grading an assignment involves looking into writing style as well. The structure of sentences, diversity of vocabulary, and sophistication of the writing style are considered when allocating marks, and the writing style has an enormous impact on the readability of a document. Therefore readability is included as a dimension to cover that aspect. Flesch Reading Ease (Flesch, 1948) and Flesch-Kincaid Grade Level (Kincaid, Fishburne Jr, Rogers, & Chissom, 1975) are very popular readability measurements, and the analysis incorporates those two measures to calculate readability scores.

For most assignments from the disciplines selected for the project, students are required to conduct literature reviews, critically evaluate situations, and support arguments and conclusions with evidence gathered from relevant literature. When referring to their research it is essential to provide citations (in-text references) and a list of references or bibliography at the end of the assignment. Therefore, reference information plays a vital role during the analysis process. The NGR software calculates the number of correctly formatted citations, number of all citations disregarding the format, the number of distinct authors in the citations, and the number of references in the reference list or bibliography. In most instances, assignment submission guidelines include a preferred referencing style. Therefore the format of the citation is checked to verify whether the correct referencing style is followed. Also, the number of distinct authors helps to identify whether the student has conducted a comprehensive literature search.

3.4 Tools and Technologies used for the Implementation

For the development of the software both C#.NET (Wagner, 2015) and Java (Oracle, 2017) programming languages have been used. The analytical component is implemented using C#.NET, whereas the user interface is developed using Java. For the user interface development, Kibana Dashboard (Elasticsearch, 2017) has been utilised. Also, the software uses functionalities provided by the Microsoft Developer Network (MSDN) when calculating the assignment statistics, proofreading errors and readability (Microsoft Developer Network, 2017a, 2017b). The Termine web tool (Frantzi et al., 2000) has been utilised to capture phrases from the assignments which are useful when identifying research, critical thinking, and discipline terms.

4 Analysis and Results

4.1 Pre-processing and Sampling

For initial analysis, it is necessary to pre-process and clean the assignment files. This includes de-identifying the assignments and cleaning the content by removing cover pages, page numbers, table of contents, figures, and other irrelevant characters and symbols. Samples were selected from the two assignments for initial analysis. For the BA subject 77 assignments were selected, and for the MKT subject 60 assignments were selected. The remaining assignments were excluded so that they could be used as a test sample to validate the tool's functionality once the software implementation was completed.

4.2 Analysis

During the analysis, values for the key evaluation dimensions are calculated for the selected samples from the two assignments. Also, the average value for each dimension is calculated, and the distribution of the average values across grades was investigated to discover relationships between those dimensions and the grades. The main objective of the analysis was to identify relationships between grades and evaluation dimensions, as that information can be utilised when generating feedback.

Initially, the selected assignments were pre-processed and directed to the NGR software to calculate assignment statistics and proof-reading errors. References and appendices were excluded from the calculation. When calculating the research, critical thinking, and discipline terms both words and phrases are extracted from the assignments to calculate the frequencies. In addition provided word lists for research, critical thinking, and discipline terms were stemmed, as stemmed words represent the root form and provide a more generalised view of those terms. For stemming, Porter's stemming algorithm (Porter, 1980) was used. Frequencies were calculated for both stemmed and non-stemmed words and phrases, and the calculation excluded references and appendices. For assignment readability, Flesch Reading Ease and Flesch-Kincaid Grade Level were calculated. Finally, referencing information was calculated for the selected assignments using the software.

4.3 Results and Discussion

Table 1 and 2 present a summary of the results calculated using the selected samples for the BA and MKT subjects.

Human marked Grade	**0-49 (F/<50%)**	**50-59 (D/ 50-59%)**	**60-69 (C/ 60-69%)**	**70-79 (B/ 70-79%)**	**80-100 (A/80+%)**	**All**
# of Assignments	10	14	17	26	10	77
Critical Thinking, Research, & Discipline Terms						
Av # of Critical Thinking Words	11.4	10.2	10.9	12.0	12.9	11.5
Av # of Research Words	19.7	20.6	22.9	23.9	24.8	22.7
Av # of Discipline Words	54.9	64.4	65.8	70.8	91.1	69.1
Citations & References						
Av Citations	11.0	9.3	8.4	14.6	13.6	11.7
Av Distinct Authors	2.8	5.0	3.0	5.2	5.1	4.4
Av # of References	4.9	7.9	7.5	8.9	8.4	7.8
Word Statistics & Proofreading Errors						
Av Word Count	1445	1503	1650	1726	1765	1637
Av Paragraph Count	7	10	6	7	9	8
Av Spelling Error per 1000 Words	5.6	6.0	6.6	7.9	7.7	6.9
Av Grammar Error per 1000 Words	2.5	1.3	1.0	1.3	0.6	1.3
Readability						
Flesch Reading Ease	44.1	41.4	38.4	33.3	28.2	36.6
Flesch-Kincaid Grade Level	13.5	13.6	14.5	15.6	16.2	14.8

Table 1: Summary results for BA subject

Human marked Grade	**0-19 (F/<50%)**	**20-23 (D/ 50-59%)**	**24-27 (C/ 60-69%)**	**28-31 (B/ 70-79%)**	**32-40 (A/80+%)**	**All**
# of Assignments	3	12	18	20	7	60
Critical Thinking, Research, & Discipline Terms						
Av # Critical Thinking terms	16.7	17.5	15.7	17.7	18.7	16.3
Av # of Research Words	41.3	49.4	40.1	42.5	40.6	42.9
Av # of Discipline Words	49.3	55.3	49.7	46.0	87.6	54.0
Citations & References						
Av Citations	9.7	17.7	12	13.2	30.7	15.2
Av Distinct Authors	5.3	8.25	7.2	6.8	13.6	7.9
Av # of References	9.7	12.1	12.4	12.6	18.6	13.0
Word Statistics & Proofreading Errors						
Av Word Count	2151	2603	2697	2758	3284	2739
Av Paragraph Count	45	66	66	68	110	70
Av Spelling Error per 1000 Words	4.3	10.3	10.3	9.6	9.6	9.7
Av Grammar Error per 1000 Words	0.6	2.5	3.8	2.7	1.3	2.7
Readability						
Flesch Reading Ease	35.3	39.3	39.3	38.7	33.4	37.8
Flesch-Kincaid Grade Level	14.5	12.8	12.7	12.9	14.1	13.0

Table 2: Summary Results for MKT subject

The two assignments selected for analysis belong to different disciplines, and degree levels. This meant each had quite different expectations of the students. The BA assignment was an essay on a themed question which expected students to refer to literature, present arguments and draw conclusions. However, the MKT assignment was focused on preparing an executive report on the marketing analysis of a company. The initial analysis results conveyed the differences between the two assignments. As an example, the average number of paragraphs for marketing assignments was very high when compared with the BA assignment, as content was presented as a report with bullet points or short paragraphs.

The analysis results also conveyed that the readability values had a significant relationship with the assignment marks, particularly for the BA subject. According to the Flesch Reading Ease score, when the readability score decreases, it indicates that the document is more suited to an audience with a higher education. This relationship seems more prominent for the essay-type assignments based on findings from the analysis. Figure 2 presents the relationship between the Flesch Reading Ease score and assignment marks for the BA subject.

According to above analysis, it can be concluded that different types of assignments are made up of different characteristics. Therefore, it is vital to consider the assignment type when designing the assessment criteria. The requirement for customising the evaluation measures based on the assignment type is important for obtaining accurate analysis outcomes and therefore will be incorporated into the software by introducing a weighting mechanism for each evaluation measure. The academic staff can prioritise each evaluation measure by assigning a weight which provides the opportunity to conduct a more customised evaluation for each assignment.

Figure 3 and 4 present the assignment evaluation page and the dashboard for academic staff respectively. Similarly, students will have the opportunity to view their values for evaluation measures for assignments they have submitted, as well as suggestions for improving those assignments.

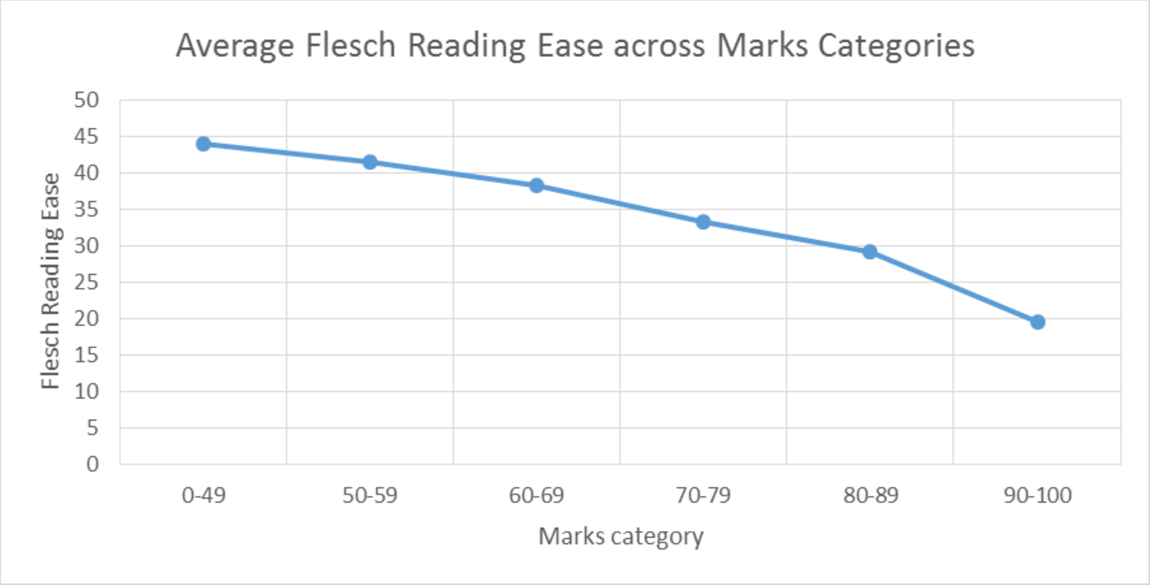
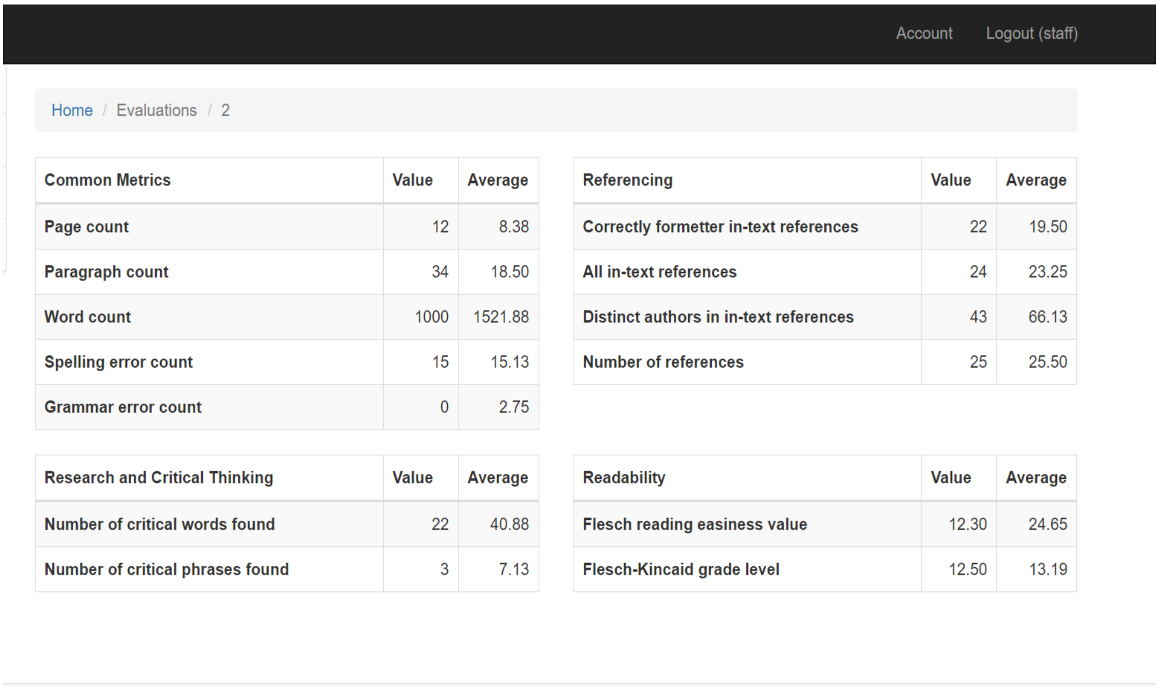


Figure 2: Average Flesch Reading Ease across marks categories for BA subject



Common Metrics	Value	Average
Page count	12	8.38
Paragraph count	34	18.50
Word count	1000	1521.88
Spelling error count	15	15.13
Grammar error count	0	2.75

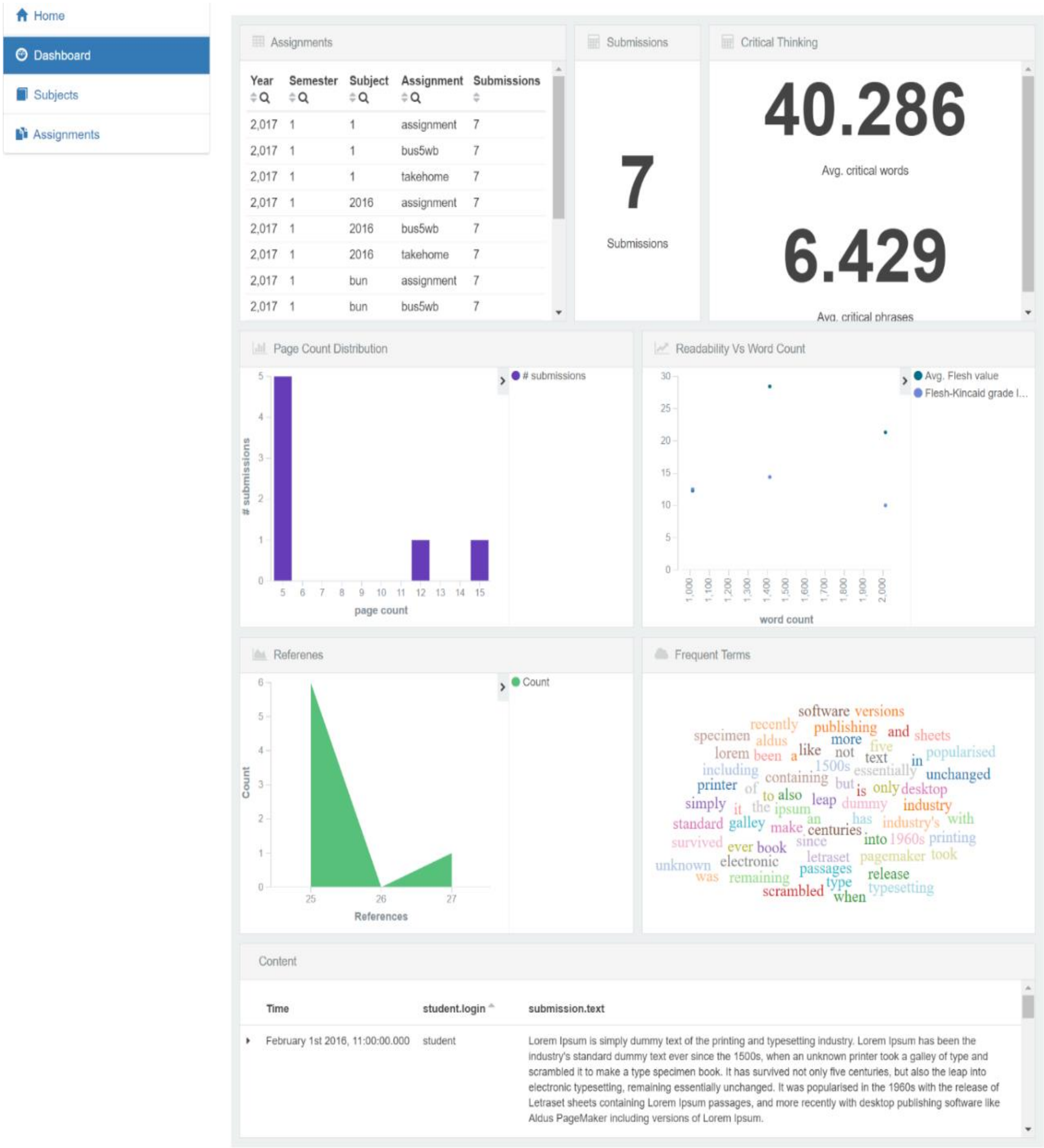
Referencing	Value	Average
Correctly formetter in-text references	22	19.50
All in-text references	24	23.25
Distinct authors in in-text references	43	66.13
Number of references	25	25.50

Research and Critical Thinking	Value	Average
Number of critical words found	22	40.88
Number of critical phrases found	3	7.13

Readability	Value	Average
Flesch reading easiness value	12.30	24.65
Flesch-Kincaid grade level	12.50	13.19

Figure 3: Assignment evaluation page in the software

Based on the values calculated for the evaluation measures, feedback will be generated for students. Moreover, threshold values will be maintained by the software to decide whether sufficient information is presented for each evaluation measure. As an example, a threshold for reference count and number of distinct authors can be defined. If the assignment statistics for reference count and number of distinct authors is less than the pre-defined threshold values, then feedback will be generated to inform the student that they need to conduct a comprehensive literature review by referring to articles published by different authors. Similarly, assignment statistics will be used to advise students about the issues associated with the structure of their assignments. Research, critical thinking, and discipline term counts will inform students whether they have sufficient content from each category.


Home
Dashboard
Subjects
Assignments

Assignments

Year	Semester	Subject	Assignment	Submissions
2,017	1	1	assignment	7
2,017	1	1	bus5wb	7
2,017	1	1	takehome	7
2,017	1	2016	assignment	7
2,017	1	2016	bus5wb	7
2,017	1	2016	takehome	7
2,017	1	bun	assignment	7
2,017	1	bun	bus5wb	7

Submissions

7

Submissions

Critical Thinking

40.286

Avg. critical words

6.429

Avg. critical phrases

Page Count Distribution
# submissions
5
4
3
2
1
0
5 6 7 8 9 10 11 12 13 14 15
page count
Readability Vs Word Count
Avg. Flesh value
Flesh-Kincaid grade l...
30
25
20
15
10
5
0
1,000 1,100 1,200 1,300 1,400 1,500 1,600 1,700 1,800 1,900 2,000
word count
Referenes
Count
6
5
4
3
2
1
0
25 26 27
References
Frequent Terms
software versions recently publishing and sheets specimen aldus more five lorem been a like not text in popularised including 1500s essentially unchanged printer of containing but is only desktop simply it the to also ipsum leap dummy industry standard galley make an centuries has industry's with survived ever book since into 1960s printing unknown electronic letraset pagemaker took was remaining passages release scrambled type when typesetting

Content

Time	student.login	submission.text
February 1st 2016, 11:00:00.000	student	Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum passages, and more recently with desktop publishing software like Aldus PageMaker including versions of Lorem Ipsum.

Figure 4: Dashboard for academic staff in the software

5 Further Analysis

5.1 Cluster Analysis

As previously discussed, clustering of the assignment data could provide academics with more insights into students' performance. Therefore, attribute clustering has been conducted for all the assignments from the two subjects. The attributes selected for clustering are listed below.

- Spelling errors per 1000 words
- Grammar errors per 1000 words
- Number of references
- Readability measured by Flesch Reading Ease score
- Research terms
- Critical thinking terms

- Discipline terms

SAS Enterprise Miner (SAS, 2016) was used to generate clusters. After an initial exploration of the data, it was decided to use five clusters for the extraction of segments for the BA assignments and four for the MKT assignments. The decision for the total number of clusters was made by trialling different cluster sizes with the main focus to obtain approximately five segments aligned with the five marking grades (A, B, C, D, and F). For the BA subject five segments could be obtained with a reasonable number of assignments in each segment. However, when the same procedure was executed for the MKT subject, one segment was very small when compared to the other segments. Therefore, clustering was executed to obtain only four segments for the MKT subject.

Clustering is a useful approach when looking for similarities across the assignments and is also useful in informing academics how assignments have similar characteristics which can be used when providing feedback to students and evaluating the performance of students on assignments.

5.2 Discussion

Based on the seven dimensions used in the study, SAS Enterprise Miner has generated segment plots and segment profiles for the two subjects. The segment plots are presented in Figures 5 and 6 for the BA and MKT subjects respectively.

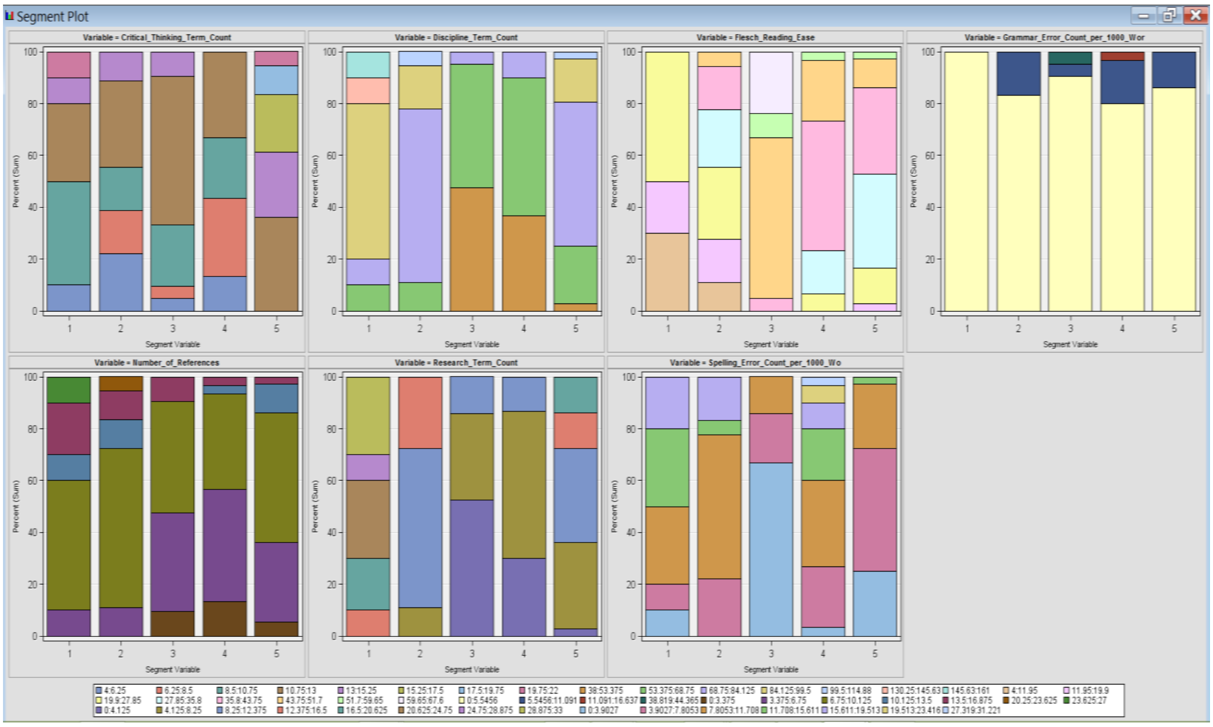


Figure 5: Segment Plot for BA Subject

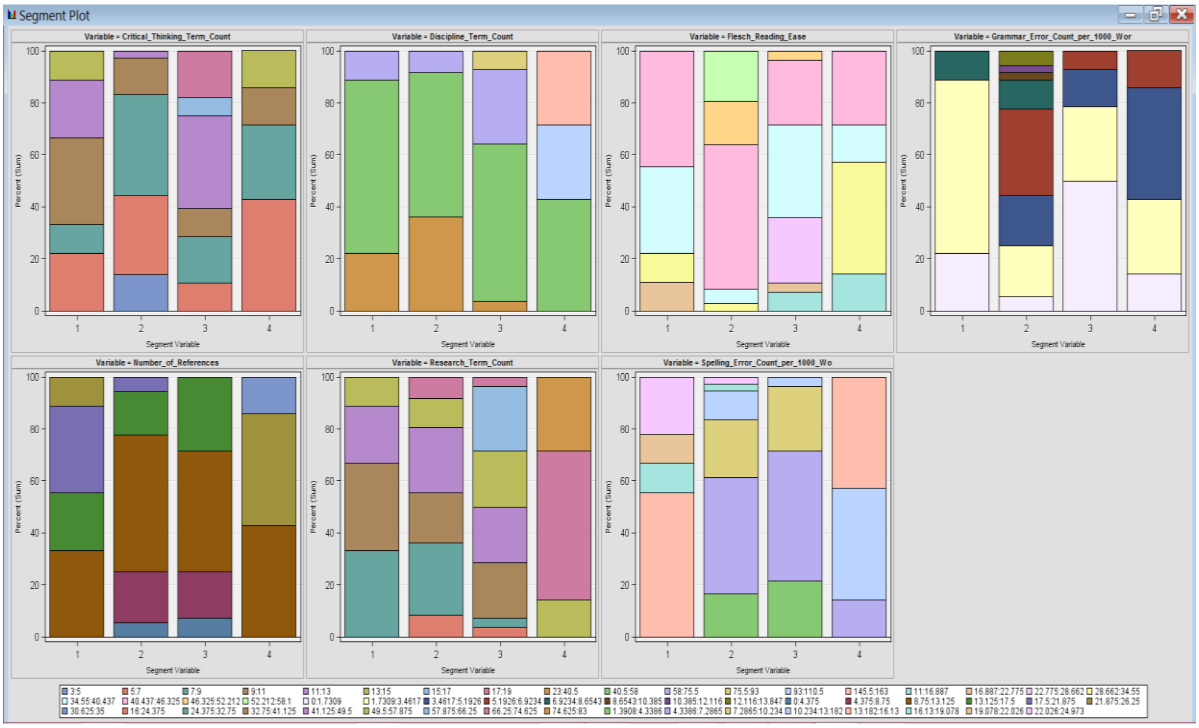


Figure 6: Segment Plot for MKT subject

An examination of the segment plots highlighted the performance of each dimension for each segment and, in association with the segment profiles (not presented here) allowed for a picture of segments to be revealed. A summary of the key features of each segment is provided in Tables 3 and 4 below for BA and MKT subjects respectively.

Segment	**Segment description**
1 (n=10)	High on research and discipline terms, low on readability
2 (n=18)	Average performance across the discipline and research terms with low readability and critical thinking terms
3 (n=21)	High readability with low performance on the discipline and research terms
4 (n=30)	Low discipline and research terms with moderate readability, and referencing
5 (n=36)	High critical thinking terms with moderate readability, discipline and research terms. Low spelling and grammar errors

Table 3: A summary of key features of each segment for the BA subject

Segment	**Segment description**
1 (n=9)	High spelling errors and moderate discipline terms, readability, critical thinking and referencing
2 (n=36)	High readability with remaining items low or moderate
3 (n=28)	Low grammar errors but high critical thinking and research terms
4 (n=7)	High on all dimensions except for grammar errors and critical thinking

Table 4: A summary of key features of each segment for the MKT subject

Five segments have been identified in the clustering results for the BA subject. Segment one consisted of high research and discipline terms and low readability, indicating stronger writing skills. Segment two had mainly moderate or low measures across the descriptors, while segment three had a high readability score with low discipline term/phrase use, research and high spelling errors, suggesting an overall poor performance. Segment four also exhibited overall poor performance with low discipline and research terms, with moderate readability and referencing. High critical thinking scores with moderate readability, discipline and research terms were noticeable in segment five, where spelling and grammar errors were low.

Therefore, a good analysis by students but only average coverage of key concepts in the assignment was evident in segment five. According to the cluster analysis results, segment one was the best performing cluster, and when compared with the grades allocated by academics, it was observed that all the assignments belonging to segment one were awarded higher marks by human markers (marks ranged from 76% to 90%). When the allocated marks were considered for each other segment, it was observed that those segments consisted of a mixture of low, moderate, and high performing assignments. In particular, segment four, which exhibited overall poor performance, had the highest number of fail grades.

Four segments were identified in the cluster results for the MKT subject. High spelling error rate and moderate to low scores on the remaining dimensions were observed in segment one. Segment two was the largest sized group and had a high readability score, but was low on most of the remaining dimensions suggesting poorer performing assignments. Segment three also consisted of a large sample which was characterised by low grammar errors and high critical thinking and research terms. Higher values for all dimensions except for grammar errors and critical thinking were noticeable in segment four. High performance in most of the dimensions and a low grammar error score suggested better performance by students. When examined in relation to the marks allocated by academics, it was observed that most of the assignments belonging to this segment had higher marks, where four out of the seven assignments had marks equal to or greater than 77.5%. For each other segment, it was observed that those segments contained a mixture of low, moderate, and high performing assignments.

The results provide useful insights to academics, explaining how rubric elements are performed across the cohort. For example, in the BA subject, the largest segment performed well on critical thinking but only moderately on readability and discussion of discipline and research terms. In contrast, better overall performance was identified for segment one, especially regarding research and discipline terms and readability, although the use of critical thinking terms was only moderate. Furthermore, weaker writing skills were exhibited in segment three due to higher readability scores combined with infrequent use of discipline and research terms. Subject coordinators can learn from these findings, identifying areas of their curriculum that require revision and identifying students who require additional support in key areas of the curriculum.

A similar approach undertaken in the MKT subject revealed groupings of assignments that performed highly across multiple dimensions (segment four) and poorly across most of the dimensions (segment two). Of concern in this latter segment was a significant number of students doing poorly in comparison to the remaining cohort.

While the research reported here has highlighted an approach to assignment analysis, of equal importance in this area has been developing an understanding of what is expected in an assignment by the subject coordinator and markers. In undertaking this work, it has been observed that markers do not take a structured approach to assessment marking, and provide inconsistent feedback to the subject coordinator on the shortcomings of an assignment. Furthermore, it is often difficult for the subject coordinator to grasp a clear overview of performance across the whole subject, particularly because of the individual biases of different markers. The approach being developed in the NGR project provides much greater insight into the performance of students at the macro level because of its consistency.

Of interest in this analysis was the performance of each subject across multiple dimensions. By looking at how each subject performed on each dimension, insights can be provided to academics about large classes that assist them to better moderate assessment, better scaffold assessment skills acquisition in class, and gain a better understanding of student performance overall. As the cluster analysis facilitates valuable information about assignments, clustering functionality will be incorporated next into the software to provide more advanced analysis and efficient feedback generation.

6 Future Directions

6.1 Challenges and Limitations

Many challenges are faced by the NGR project while assessing student assignments. One of the key challenges is conducting the analysis when the assignment does not comply with submission guidelines. A common mistake by students during assignment submission is that they only pay attention to the main question, ignoring other important factors such as how the assignment should be formatted or how references should be presented. As an example, if references are included as footnotes in the assignment it becomes difficult to calculate referencing information accurately. However, as mentioned in the methodology section, the assignment analysis process is designed to be aligned with the rubric and submission guidelines. In situations where assignments do not comply with the submission guidelines, it leads to difficulties conducting analysis and obtaining accurate results. As such it was considered a limitation in the current version of the software.

Another major challenge is identifying research and discipline terms. Currently, this process is done manually by academics and thus inefficient. Moreover, measuring assignment coverage for research and discipline concepts based on a term list is not efficient and requires enhancements. One possible solution is to create an ontology or refer to an existing one to obtain a diverse set of terms related to a subject discipline. Such an approach would be highly advantageous to identify whether assignments have addressed topics related to the subject discipline. In addition, the software possesses limitations in measuring referencing information as well, as it requires the citations and references to follow an acceptable style. Moreover, limiting the evaluation to a set of pre-defined evaluation measures is also a limitation associated with the NGR software. If a new evaluation measure needs to be incorporated, then the software needs to be modified accordingly. As future work above mentioned, limitations can be addressed to enhance the usability of the NGR software and the accuracy of the analysis results.

6.2 Future Work

As the current analysis techniques only facilitate the primary content analysis of the assignments, an investigation has been undertaken to identify further analytical methods which might be advantageous and integrated into the software. Relevant techniques are listed below.

- Advanced style, spelling and grammar check
- Topic mining
- Capturing lexical chains
- Discriminant analysis

The current version of the software captures spelling and grammar errors, however, incorporating advanced style, spelling and grammar checks could provide valuable insights into the analysis. Such analysis allows capturing redundant phrases, non-standard phrases, commonly confused words, misuse of capitalisation and punctuation errors. Moreover, differentiation between Australian, American and British English can be conducted to verify whether the assignment complies with the preferred language. LanguageTool (LanguageTool, 2017) is an open source proofreading software which facilitates advanced style, spell and grammar checks. Therefore, in the future, functionality provided by LanguageTool will be integrated to achieve a more comprehensive analysis.

Another important analysis technique is topic mining, which allows the discovery of major topics associated with each assignment. Discovering topics is important as it identifies the key concepts covered in the assignments and theories referred from textbooks. Several topic mining software tools are currently available, including Mallet (McCallum, 2002) a popular open source tool which facilitates topic mining via Latent Dirichlet Allocation (LDA) (David M Blei, Ng, & Jordan, 2003), Pachinko Allocation (Li & McCallum, 2006), and Hierarchical LDA

(David M. Blei, Jordan, Griffiths, & Tenenbaum, 2003) algorithms. The software captures topics with topical terms and topic composition for the documents. In the future, the applicability of topic mining functionality will be investigated and if successful, integrated into the current version of the software using Mallet.

Lexical chains allow the capture of sequences of related words in a document and considered a useful mechanism to capture semantic content of a document. Lexical chains have been utilised for document summarisation tasks, as they capture the main themes discussed in the documents (Wei, Lu, Chang, Zhou, & Bao, 2015). Therefore, the possibility of integrating lexical chains into the analysis process will be investigated. Capturing lexical chains will be beneficial in measuring whether assignments cover the necessary discipline concepts and can be utilised in grading and the feedback generation process.

Discriminant analysis is a popular classification technique and consists of different variations that include linear and quadratic forms. This technique can be used to classify assignments into grades based on key assignment characteristics and useful for assignment evaluation and feedback generation. However, this technique requires a set of already marked assignments for training, as it learns characteristics of each grade using those samples. Therefore, the successful implementation of discriminant analysis will depend on the availability of training samples and how each grade can be clearly distinguished based on distinct characteristics of the assignments.

Integrating the above techniques will be beneficial for a more comprehensive analysis of the assignments and will be considered for future work on the project.

7 Conclusion

The results presented in this paper highlight opportunities for academics, and in particular subject coordinators of large classes, to undertake automated analysis of assignments in order to obtain greater understanding of student performance in major pieces of assessment against marking rubrics using tools from learning analytics. Work on the project is continuing with the preparation of software to automate much of the feedback process, and to present data in a user-friendly dashboard for subject coordinators that can also be used to provide feedback to students on individual performance in their assignments. This would include the opportunity for subject coordinators to drill down into the data to identify students at-risk. Further work to refine the accuracy of the tool could include the use of additional measures, including behavioural measures.

Acknowledgments

The authors gratefully acknowledge support from a Digital Learning Strategy Innovation Grant provided through La Trobe Learning and Teaching at La Trobe University, Melbourne. They would also like to acknowledge the assistance of James Heath in research assistance associated with this work.

References

1st International Conference on Learning Analytics and Knowledge. (2010). LAK '11: 1st International Conference on Learning Analytics and Knowledge 2011. Retrieved from https://tekri.athabascau.ca/analytics/

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*: O'Reilly Media, Inc.

Blei, D. M., Jordan, M. I., Griffiths, T. L., & Tenenbaum, J. B. (2003). *Hierarchical topic models and the nested chinese restaurant process*. Paper presented at the Proceedings of the 16th International Conference on Neural Information Processing Systems, Whistler, British Columbia, Canada.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine learning research, 3*(Jan), 993-1022. Retrieved from http://www.jmlr.org/papers/v3/

Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology, 37*(1), 51-89. doi:10.1002/aris.1440370103

Coh-Metrix. (n.d.). Coh-Metrix. Retrieved from http://cohmetrix.com/

Dawson, P. (2017). Assessment rubrics: towards clearer and more replicable design, research and practice. *Assessment & Evaluation in Higher Education, 42*(3), 347-360. doi:10.1080/02602938.2015.1111294

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*(1), 7-24. doi:https://doi.org/10.1016/j.asw.2012.10.002

Educational Data Mining. (2017). Retrieved from International Educational Data Mining Society website: http://www.educationaldatamining.org/

Elasticsearch. (2017). Kibana. Retrieved from https://www.elastic.co/products/kibana

ETS®. (2017a). Criterion®. Retrieved from https://www.ets.org/criterion

ETS®. (2017b). TOEFL: For Academic Institutions: Scores. Retrieved from https://www.ets.org/toefl/institutions/scores

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology, 32*(3), 221-233. doi:http://dx.doi.org/10.1037/h0057532

Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi-word terms:. the C-value/NC-value method. *International Journal on Digital Libraries, 3*(2), 115-130. doi:10.1007/s007999900023

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel* (No. RBR-8-75). Retrieved from Defense Technical Information Center website: http://www.dtic.mil/get-tr-doc/pdf?AD=ADA006655

LanguageTool. (2017). LanguageTool Style and Grammar Check. Retrieved from https://www.languagetool.org/

Li, W., & McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. *Proceedings of the 23rd international conference on Machine learning*, 577-584. doi:10.1145/1143844.1143917

Liddy, E. D. (2001). Natural language processing. In M. Drake (Ed.), *Encyclopedia of Library and Information Science* (2nd ed., pp. 2126-2136). New York, NY: Marcel Decker, Inc.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55-60. doi:10.3115/v1/P14-5010

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. Retrieved from http://mallet.cs.umass.edu

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing, 23*(Supplement C), 35-59. doi:https://doi.org/10.1016/j.asw.2014.09.002

Microsoft Developer Network. (2017a). Document.ComputeStatistics Method (WdStatistic, Object) (Microsoft.Office.Tools.Word). Retrieved from https://msdn.microsoft.com/en-us/library/microsoft.office.tools.word.document.computestatistics.aspx

Microsoft Developer Network. (2017b). ProofreadingErrors interface (Microsoft.Office.Interop.Word). Retrieved from https://msdn.microsoft.com/en-us/library/microsoft.office.interop.word.proofreadingerrors.aspx

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM, 38*(11), 39-41. doi:10.1145/219717.219748

Oracle. (2017). Oracle Technology Network for Java Developers | Oracle Technology Network | Oracle. Retrieved from http://www.oracle.com/technetwork/java/index.html

Page, E. B. (1966). The Imminence of... Grading Essays by Computer. *The Phi Delta Kappan, 47*(5), 238-243. Retrieved from http://www.jstor.org/stable/20371545

Paul, R. (1995). *Critical thinking: How to prepare students for a rapidly changing world*: Foundation for Critical Thinking.

Pearson. (2012). Objective fact sheet. Retrieved from https://pearsonpte.com/wp-content/uploads/2014/07/ObjectiveFactsheet.pdf

Pearson. (2017). WriteToLearn™. Retrieved from https://www.pearsonassessments.com/products/100000030/writetolearn.html#tab-details

Popham, W. J. (1997). What's wrong-and what's right-with rubrics. *Educational leadership, 55*(2), 72-75. Retrieved from http://www.ascd.org/publications/educational-leadership/oct97/vol55/num02/What's-Wrong%E2%80%94and-What's-Right%E2%80%94with-Rubrics.aspx

Porter, M. F. (1980). An algorithm for suffix stripping. *Program-Automated Library and Information Systems, 14*(3), 130-137. doi:10.1108/eb046814

SAS. (2016). Data Mining Software, Model Development and Deployment, SAS Enterprise Miner | SAS. Retrieved from https://www.sas.com/en_us/software/enterprise-miner.html

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation: Current applications and new directions*. New York, NY: Routledge.

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Siemens, G., & Baker, R. S. J. d. (2012). Learning analytics and educational data mining: towards communication and collaboration. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 252-254. doi:10.1145/2330601.2330661

Turnitin. (2017). Turnitin - What We Offer : Revision Assistant. Retrieved from http://www.turnitin.com/en_us/what-we-offer/revision-assistant

Vantage Learning. (2016). MY Access!® Writing and Assessment Solution | Vantage Learning. Retrieved from http://www.vantagelearning.com/products/my-access-school-edition/

Vitartas, P., Ahmed, T., Alahakoon, D., Midford, S., Nathawitharana, N., Ong, K. L., & Sullivan-Mort, G. (2016, 5-7 December). *Using learning analytics to guide learning: an analysis of marketing assignments*. Paper presented at the Proceedings of the Australian and New Zealand Marketing Academy Conference (ANZMAC), Christchurch, New Zealand.

Wagner, B. (2015). Introduction to the C# Language and the .NET Framework. Retrieved from https://docs.microsoft.com/en-us/dotnet/csharp/getting-started/introduction-to-the-csharp-language-and-the-net-framework

Wei, T., Lu, Y., Chang, H., Zhou, Q., & Bao, X. (2015). A semantic approach for text clustering using WordNet and lexical chains. *Expert Systems with Applications, 42*(4), 2264-2275. doi:https://doi.org/10.1016/j.eswa.2014.10.023

White, B., & Larusson, J. A. (2010). Detecting the "point of originality" in student writing. *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, 817-820. doi:10.1145/1868914.1869037

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

WriteLab. (n.d.). WriteLab Homepage. Retrieved from http://www.writelab.com