MONASH University

FIT5210

Semester 1, 2023
Minor Thesis
Final Thesis

Author Name: Siddharth Gupte

Thesis Title: Facilitating Sensemaking in Learner-centric Feedback with NLP

Student ID: 31524923

Course: Master of Data Science (C6004)

Supervisor: Yi-Shan Tsai

Co-Supervisor: Guanliang Chen

Version: 4.0

Reference Style: APA 7

Word Count: 7980

Date Created: 01/04/2023

Last Modified: 16/06/2023

Table of Contents

Acknowledgements

Let me start by thanking my supervisors Yi-Shan Tsai and Guanliang Chen for giving me the opportunity to work on this minor thesis project. Their guidance throughout the course of this thesis facilitated my entry into the world of feedback research and its numerous terminologies. With their help, I was able to form the research questions to be answered by this project. Furthermore, a special thanks to Ahmad Aldino (ahmad.aldino@monash.edu) for dividing the task of labelling the feedback samples with me and for writing the initial comment examples for each rubric. I also appreciate his help in introducing me to the Yedda Text Annotation Tool that we customised for this study. Furthermore, some of his coding-based inputs for RQ1, RQ2 and RQ3 helped me enrich the findings of this thesis.

Abstract

Feedback is crucial for teaching in higher education institutions. To enhance the effectiveness of feedback, it must be learner-centred. Learner-centred feedback ensures that the learner takes an active role in the feedback process by providing context, actionable information, and positive reinforcement. While several studies have contributed towards understanding learner-centred feedback, it is unclear how current feedback practices in higher educational institutions align with these principles. This alignment must be studied to measure feedback quality and improve it accordingly. Thus, we used a learner-centred framework as our guide to study the feedback pieces and understand them from a sensemaking point of view. To do this, we collected and labelled assessment feedback data from Semester 1, 2021 of the Master of Data Science course at Monash University. Using NLP techniques, we found important details of the sensemaking component including its linguistic complexities, its co-occurrence with other components, the student score, and its alignment with Bloom's Taxonomy. Finally, we extended this study by devising NLP classification models to identify Sensemaking comments in the feedback. They achieved exceptional accuracy in predicting the sensemaking aspect.

Keywords

Sensemaking, Future Impact, Learner-centred, Classification, NLP, Machine Learning, DistilBERT

1. Introduction

Feedback is information that is given to students and used by them with respect to their academic progress to improve their work and overall learning (Ryan et al., 2022). The appropriate construction of feedback can potentially cause an effective change in the learning of the student.

Feedback must help convert the student's current understanding to the desired understanding of the subject matter (Hattie & Timperley, 2007). When it achieves this purpose, it is said to be learner-centred, efficiently supporting students in their learning. Learner-centred feedback enables the learner and the education provider to measure the success and failure of learning techniques (Ryan et al., 2021).

Learner-centred feedback has been of great interest in academic literature. With the turn of the last decade, there was a shift in feedback studies where students were brought into a more prominent role in their feedback process, beyond their post-feedback performance (Dawson et al., 2019).

Despite this, there are three problems in feedback literature:

Conceptual Limitations

Several studies still rely on Hattie and Timperley's feedback model (Ryan et al., 2021) which is more teacher-oriented than student-oriented (Mandouit & Hattie, 2023). Scholars have argued that feedback practices must move away from that model to effectively align with learner-centred feedback principles (Ryan et al., 2021). Additionally, studies that have introduced frameworks such as feedback triangle (Yang & Carless, 2013) and the components of effective learner-centred feedback (Ryan et al., 2021) are relatively new. The adoption of these frameworks in feedback practices is unclear. Moreover, studies in the latter part of the last decade have focused on a generalised approach to feedback that mostly includes student performance improvement (Dawson et al., 2019).

These issues indicate a gap between the theoretical understanding of feedback in recent studies and current feedback practices. There is a need to bridge this gap and find a way to measure feedback quality in a tangible manner.

Operational Obstacles

The amount of natural text to be processed has significantly increased in recent times. It has introduced the following issues in adequately identifying feedback-centric components. Analysing a large amount of text is time-consuming for a human being (Katz et al., 2021). Furthermore, the limitation of the human brain may introduce consistency issues while analysing text over a long period of time (Katz et al., 2021). Machine learning and Natural language processing techniques have the capability to foster effective feedback practices due to their predictive and classification capabilities (Shaik et al., 2022). Well-designed NLP models can automate the process of analysing feedback and provide useful results (Shaik et al., 2022). However, their full potential has not been unleashed in feedback analysis (Lin et al., 2023).

Sensemaking Literature Scarcity

As proposed by Ryan et al. (Ryan et al., 2021), the sensemaking component of learner-centred feedback deals with communicating the strengths and weaknesses of the student's work. It is the primary foundation of good feedback as it provides awareness on the specifics of the student's work. If the student is not able to understand and interpret the feedback, its future impact will be diminished (Ryan et al., 2022). However, Sensemaking has not been addressed sufficiently in previous studies (Ryan et al., 2022).

Therefore, this project focused on facilitating sensemaking (Ryan et al., 2021) with NLP techniques. It will contribute towards a larger study that investigates all three components of learner-centred feedback (Ryan et al., 2021). Thus, we concocted the following research questions for this study:

RQ1 To what extent does current feedback practice align with the Sensemaking aspect of learner-centred feedback?

RQ2 To what extent does the presence of the Sensemaking aspect in learner-centred feedback vary by student performance?

RQ3 How does the sensemaking aspect of learner-centred feedback align with Bloom's Taxonomy of Action Verbs?

RQ4 To what extent can machine learning/NLP models correctly identify the sensemaking aspect of feedback?

To answer these questions, we randomly selected feedbacks and labelled them using eight learner-centred comment types.

For **RQ1**, we used NLP techniques to understand different aspects of the Sensemaking comments such as their textual complexities and reading quality.

For **RQ2**, we performed inferential statistics and student grade distribution analysis to understand the dependency of Sensemaking on the student score and compared it to the dependency of its multi-label combination with other components with the score.

For **RQ3**, we used NLP to extract action verbs from the feedback text and match them to the verbs that fall under each level of Bloom's Taxonomy. When a match was found, the corresponding levels were assigned to that piece of feedback.

For **RQ4**, we compared the performance of different classification models in identifying the sensemaking aspect in the feedback. Furthermore, we compared two NLP feature extraction methods TF-IDF (Nasim et al., 2017) and LIWC (Boyd et al., 2022) to enrich the results of the models.

2. Background

Several studies have explored the importance of feedback in students' learning process in higher education (Dawson et al., 2019; Henderson et al., 2019; Lim et al., 2021; Ryan et al., 2021; Tsai et al., 2021). Understanding the frameworks described in these studies is crucial in the understanding of learner-centred feedback components (Tsai et al., 2021).

Whilst perusing existing feedback studies, we aimed to review four different areas of interest.

2.1 The Feedback Gap

Despite the importance of feedback being agreed upon across the world (Hattie, 2008; Hattie & Timperley, 2007; Brooks et al., 2021), research suggests that it has not been utilised to its full potential in feedback practice (Hattie et al., 2016; Brooks et al., 2021). For example, it was observed that the students' understanding of their learning outcomes and how to improve them (Lee et al., 2019; Timperley & Parr, 2009; Brooks et al., 2021) is given a (Wiliam, 2017; Brooks et al., 2021) formalistic coverage at best without tackling it in depth.

The 'feedback gap' is a term that has been used to describe this disparateness between the understanding of effective feedback in research and deploying it in practice (Dawson et al., 2018; Evans, 2013; Iraj et al., 2021). Thus, analysing this gap will help explain the extent to which it is effective in student learning (Iraj et al., 2021).

2.2 A Traditional Framework

For a long time, the Hattie and Timperley feedback model (Hattie & Timperley, 2007) has dictated feedback literacy. Overall, this model (Hattie & Timperley, 2007) introduced three feedback characteristics: (i) Learning goals (ii) Current status (iii) Next course of action. These scenarios traverse four levels: (i) Self focus, (ii) Task, (iii) Process, and (iv) Self-regulation (Lin et al., 2023; Hattie & Timperley, 2007). The same study (Hattie & Timperley, 2007) found that the feedback had a transformational effect on the student's improvement when it focused more on the student's positives rather than negatives (Hattie & Timperley, 2007). Thus, it (Hattie & Timperley, 2007) mainly focuses on student achievement (Ryan et al., 2021).

Subsequently, it was observed that for feedback to be effective, it must be of a quality that encourages consistent student engagement (Boud and Molloy 2013; Henderson et al., 2019; Boud & Dawson, 2023). It must also be contextually sound to be impactful (Ryan et al., 2021).

A study likened the feedback model (Hattie & Timperley, 2007) to the idea of an engineering model that downplays the importance of the learner (Boud & Molloy, 2013).

2.3 Learner Centred Feedback

Following the concoction of the feedback model (Hattie & Timperley, 2007), an increased interest in feedback literature has led to a conceptual shift in the feedback paradigm (Dawson et al. 2019; Ryan et al., 2021).

Several new frameworks have been introduced with a focus on learner-centred feedback. For example, the feedback triangle was introduced to enhance feedback dialogue (Yang & Carless, 2013). Another study proposed a more sustainable model, Mark 2 that acknowledges the active role of the learner (Boud & Molloy, 2013; Winstone et al., 2022). Moreover, in a revisitation of the feedback model (Hattie & Timperley, 2007), an improved version that considers the perspective of the learner was proposed (Mandouit & Hattie, 2023).

To capture learner-centred feedback, Ryan et al. (Ryan et al., 2021) undertook a systematic three-phase effort to develop a taxonomy of feedback components (Ryan et al., 2021). A three-component framework was devised from this taxonomy:

- Enable Sensemaking: Provide the student with contextualised information about their work.
- Focus on Future Impact: Give actionable information for improvement.
- Support Agency: Develop a quality dialogue with students by addressing to their socio-affective needs.

Additionally, Ryan et al. (Ryan et al., 2021) also highlighted the dependency of Future Impact on succinct Sensemaking comments. Thus, we also involved Future Impact and Support Agency in our experiments, with respect to their co-dependency on Sensemaking (Ryan et al., 2021) to help us understand Sensemaking better.

Language complexities can also determine the effectiveness of Sensemaking elements. Existing studies of sensemaking have argued that learners' understanding is hampered with complicated wording (Ryan et al., 2022), brief comments or an overreliance on learner agency (Dohrer, 1991; Esterhazy and Damşa, 2019; Zhang and Zheng 2018; Ryan et al., 2022). Therefore, we aim to use NLP techniques to break down the Sensemaking text and understand its language complexities.

Moreover, since Sensemaking focuses on identifying the strengths and weaknesses of the student's work, it is related to cognitive development. Another framework that is aimed towards cognitive development is Bloom's Taxonomy (Das et al., 2022).

Bloom's Taxonomy consists of six hierarchically arranged levels often used to describe the cognitive complexity of learning objectives (Li et al., 2022). The revised levels are given in table 1.

Table 1: Bloom’s Taxonomy of Action Verbs (Pickard, 2007)

Level	Description	Sample Verbs
Remembering	Demonstrate perfect recall of previously learned items.	Choose, Find, Select, Show
Understanding	Show a complex understanding of the subject matter.	Compare, Contrast, Interpret, Show
Applying	Can apply knowledge in new scenarios.	Apply, Identify, Organise, Develop
Analyzing	Break down concepts and identify patterns.	Analyze, Categorise, Inspect, Survey
Evaluating	Present solid arguments to defend opinions and draw criteria-based conclusions about work standards.	Appraise, Deduct, Influence, Conclude
Creating	Envision new ways of combining information or propose new ideas.	Adapt, Build, Delete, Happen

Studies have tried to map educational content with Bloom’s Taxonomy (Li et al., 2022). Li et al (Li et al., 2022) deployed models to classify learning objectives into the six levels of Bloom’s Taxonomy. Chang et al. (Chang & Chung, 2009) and Osadi et al. (Osadi et al., 2017) created systems to automatically classify teacher questions into Bloom’s Taxonomy levels.

The construction of feedback needs to consider learning outcomes in line with the ‘feed-up’ principle (Hattie & Timperley, 2007) to contextualise the strengths and weaknesses of student work.

Considering this to be an important cognitive paradigm, we aimed to align Bloom’s Taxonomy levels (Krathwohl, 2002) with Sensemaking by detecting the key verbs commonly used in these levels.

2.4 Use of Machine Learning/NLP techniques

Several strides have been made in the domain of machine learning and NLP techniques with the increase in computational power (Katz et al., 2021) in recent years. NLP techniques like Bag-of-words, Ngrams and TF-IDF have been highlighted (Botelho et al., 2023) for their capacity to capture context in the text and improve text classification performance.

Furthermore, natural language processing and deep learning have expanded the ways in which textual data can be classified. For example, a pre-trained language model called Sentence-BERT can vectorize sentences to detect patterns that may not be noticeable otherwise due to being spread out across the data (Reimers & Gurevych, 2019). Similarly, the AI community HuggingFace (Wolf et al., 2020) has developed a family of state-of-the-art algorithms to handle NLP at a deep learning level which includes BERT (Devlin et al., 2019) and DistilBERT (Sanh et al., 2020).

Educational research has also achieved some success in the use of NLP techniques. For example, an automatic grading system was developed for open-ended questions by training models on TF-IDF-extracted features (Erickson et al., 2020; Botelho et al., 2023).

Two studies have made specific developments for feedback studies. One of them used models like XGBoost and BERT to classify learning objectives designed by teachers (Li et al., 2022) into the six levels of Bloom’s Taxonomy (Pickard, 2007). Another study developed classification models like Gradient Boosted Tree and Logistic Regression to predict the student grade change between two assignments based on features extracted from the feedback (Lin et al., 2023). However, the former (Li et al., 2022) analyses learning objectives and the latter (Lin et al., 2023) studies the impact of feedback features on student grade change.

The second study (Lin et al., 2023) also concedes that student grade changes may not be sufficient to encapsulate the feedback impact. Thus, we studied the feedback irrespective of the student grade change. Additionally, we aimed to study the grade accompanying a piece of feedback and its impact on the Sensemaking comments being used in the feedback.

3. Methodology

The initial methodology evolved over the course of this study following the initial literature review. **RQ1**, **RQ2** and **RQ3** were developed after studying more literature. The scope of performing classification was deemed large enough for **RQ4**. Therefore, clustering techniques were excluded from this study. We also abandoned the use of the Spacy-Python library in favour of NLTK-Python for simplicity. Finally, Bloom’s Taxonomy was excluded from **RQ4** as it is out of the scope of this study.

3.1 Feedback Data Acquisition and Labelling

We extracted assessment feedback data from Monash University’s Moodle database. We acquired a low-risk project approval (ID: 29874) from the Monash University Human Research Ethics Committee and accessed feedbacks across two semesters. Overall, these files amounted to a total of 56087 individual feedbacks. For this study, we selected the following attributes from the data:

- stdnt_otcm_grd (Marks achieved by the student)
- assign_grd_mx (Maximum marks of the assignment)
- commenttext (The feedback text)

3.1.1 Data Acquisition

A subset of the extracted data was utilised for the current study based on a specific selection criterion. The Semester 1, 2021 data was selected due to its relative recentness. Units where less than 50% of individual assessment pieces had feedback were discarded. For the scope of this study, units from the Master of Data Science course were selected, given that it is the largest Masters course at Monash University.

3.1.2 Revisiting the Components of Learner-Centred Feedback

According to Ryan et al. (Ryan et al., 2021), there are three components of learner-centred feedback. We used the 8 comment types of these components as rubrics for labelling the data.

Furthermore, we concocted examples for each of these rubrics to serve as a guide for labelling the data as shown in table 2.

Table 2: Rubrics for Labelling Feedback

Component (Ryan et al., 2021)	**Rubric**	**Example**	**Label**
Focus on Future Impact	**1.** Provide actionable information to help the student improve aspects of similar tasks they may undertake in the future	Make sure that you connect your ideas more effectively in future work (Koşar, 2021).	Impact 1
	2. Provide actionable information to help the student achieve the learning outcomes for the subject.	Try to develop your arguments further in the next assessment task (Ryan et al., 2022).	Impact 2
	3. Provide actionable information to help the student develop learning skills, processes or strategies that could be useful across and beyond their degree.	A good strategy is to read your work aloud to pick up any errors.	Impact 3
Enable Sensemaking	**4.** Highlight strengths and weaknesses in terms of specific aspects of the student's task, such as grammar, content, structure, etc (Ryan et al., 2021).	1. You provided a concise and clear definition of the key concepts. 2. This sentence is not grammatically correct (Koşar, 2021).	Sensemaking 1
	5. Summarise the overall strengths and weaknesses of the student's performance in relation to the learning outcomes/assessment criteria.	1. Overall, fantastic work! 2. You could have done better with this task.	Sensemaking 2

Support Agency	**6.** Encourage the student to take an active role by discussing their work with the teacher or tutor, engaging in further study, or seeking help from sources other than the teacher.	1. Come and see me if you would like to discuss how you could improve your next piece of work. 2. Go and see an Academic Skills Adviser to learn some tips for how to structure an essay (Ryan et al., 2021). 3. You may find it useful to search for online videos which describe the process of photosynthesis.	Agency 1
	7. Affirm student's achievement on the completed performance and/or encourage them in future work (Koşar, 2021).	Nice effort. I can tell that you tried hard on this task (Ryan et al., 2021).	Agency 2
	8. Convey information that will strengthen the teacher's relationship with the student.	I really enjoyed reading your essay, good luck with the exam (Ryan et al., 2021)!	Agency 3

As mentioned, earlier, we are focusing on Sensemaking due to the scope of the study for learner-centred feedback. However, we labelled the data with rubrics from all three components (Ryan et al., 2021) to help differentiate Sensemaking comments from them.

3.1.3 Pre-processing and Labelling the Data

We pre-processed and labelled this data through multiple stages. We used data wrangling to parse it into a cleaned form. Additionally, we created a new column that contains the percentage of the student score accompanying each feedback.

An inter-rater reliability test was conducted. with the Cohen-Kappa (Lin et al., 2023) statistic to ensure analysis vigour. This qualitative test measures the similarity between the ratings of two people and ranges from 0 to 1 with 1 being the best score (Lin et al., 2023). We reused stratified sampling to generate 30 feedbacks randomly for each test. The test was conducted 3 times. We successfully achieved consecutively favourable scores of 0.70 and 0.83.

The final random sample of 1000 feedbacks were extracted for the main labelling activity. We used a lightweight text annotation tool called Yedda to label the data (Yang et al., 2018). We modified its base code to include the 8 rubrics described in table 4.

3.2 Machine Learning/NLP Techniques – A Review

3.2.1 Natural Language Techniques

We used different NLP techniques including Tokenization (Karttunen et al., 1996), Lemmatization (Zhang et al., 2023), Part-of-Speech Tagging (POS) (Chiche & Yitagesu, 2022;

R et al., 2022), Word Clouds and Ngrams (Sulis et al., 2022) to break down the feedback text and enrich our results.

3.2.2 Classification Models

Given that we wanted to understand the extent to which the sensemaking aspect of feedback can be differentiated from the rest, it was suitable to construct a binary classifier to achieve this result. We developed five traditional machine learning classification models, Logistic regression (Pavlyshenko, 2016), Support Vector Machine, Random Forest and XGBoost (Pavlyshenko, 2016). Of these models, Logistic Regression (Pavlyshenko, 2016) and Support Vector Machine work well with binary linearly separable data while Random Forest and XGBoost (Pavlyshenko, 2016) are ensemble tree algorithms that train multiple weak models to combine their outcomes. Additionally, given the performance of deep learning algorithms, we used a distilled version of BERT (Devlin et al., 2019) called DistilBERT (Sanh et al., 2020), developed by HuggingFace (Wolf et al., 2020).

3.3 Implementation

3.3.1 Sensemaking Breakdown

For research questions **RQ1**, **RQ2** and **RQ3**, different granular levels of the data were analysed:

- Multi-Rubric and Single Rubric Occurrences: This includes unique occurrences of a combination of rubrics for a single sentence or a piece of feedback. For example, (Sensemaking 1 & Impact 1) or (Agency 2 & Sensemaking 2). It also includes unique occurrences of each rubric.

- Total Rubric Data: This refers to all occurrences of each rubric inclusive of singular occurrences and rubric combinations.

- Component Level Data: Component level data contains the total number of times the component occurs in the data, inclusive of every occurrence of its rubrics. Component refers to Sensemaking, Future Impact and Support Agency (Ryan et al., 2021).

However, we compiled the most relevant results with respect to each research question. For the individual rubrics, Sensemaking 1 and Sensemaking 2, we included their single rubric occurrences. For a view of the total rubric data, please refer to the code in Appendex 1.

To answer **RQ1**, we wanted to understand the textual complexities of the feedback comments that had a Sensemaking (Ryan et al., 2021) aspect to understand the extent of language issues (Ryan et al., 2022) in feedback like the Automated Readability index and average word count. We used Flesch's Reading Ease (Flesch, 1948) to compute the Automated Readability index. This metric ranges from 0-100. The higher the score, the better the readability for school-level students. Refer to Appendices 3.1.

Furthermore, we aimed to observe the distribution of words that are used to form sensemaking comments. We also extracted the action verbs from the text using tokenization, lemmatization, and POS tagging, considering that they may have contextual insights about the actionability of the sensemaking comments. As discussed in Table 2, Agency 2 comments affirm and encourage the student's efforts, eg. "Well done!". Agency 3 comments convey additional information aimed at fostering stronger teacher-student relationships, eg. "All the best!". However, it is important for sensemaking comments to have a positive tone to help engage the student with the feedback. Thus, we analysed the overall sentiment of Sensemaking comments in the absence of the Agency 2 and Agency 3 rubric.

For **RQ2**, we wanted to discover the distribution of grades accompanying Sensemaking comments. We created a new student grade column using the score accompanying each feedback. The grades are based on Monash University's grading system and assigned as follows:

- N (0-49)
- P (50-59)
- C (60-69)
- D (70-79)
- HD (80-100)

We also aimed to understand the extent of the dependency of the grade on the extent of Sensemaking comments. Therefore, we considered two inferential statistical methods, Chi Squared test and One-way ANOVA (Analysis of Variance) test (Mishra et al., 2019) as they are suitable for comparing categorical variables. Since One-way ANOVA (Mishra et al., 2019) is mostly preferred for three or more independent variables, we proceeded with the Chi-Squared test (Varvara et al., 2021) for this study.

To answer **RQ3**, we developed a novel method to compute Bloom's taxonomy levels for different granularities of the sensemaking component. We used POS tagging to extract the action verbs from the sensemaking comments and match them with the action verbs of Bloom's taxonomy levels (Pickard, 2007). If a match was found, the corresponding level would be attached to the sensemaking comment that contained that verb.

3.3.2 NLP Model Implementation

To answer **RQ4**, we developed models based on four machine learning algorithms and one deep learning algorithm. We used TF-IDF (Nasim et al., 2017) and LIWC (Boyd et al., 2022) to perform feature extraction for the machine learning models. TF-IDF produces the word importance for every word in a document in the corpus (Nasim et al., 2017). LIWC is a text analysis application that puts the words of a text into psychologically meaningful categories (Tausczik & Pennebaker, 2010). Both methods help depict the linguistic characteristics of a piece of text. These characteristics can ultimately be used as features to empower the training process of a machine learning model (eg. classifying the text into different categories).

All four models were trained twice, once with each feature extraction method. Given our findings regarding the weightage of multi-rubric Sensemaking combinations with Future Impact, we proceeded to devise model scenarios for **RQ4** that considered these interactions. To answer **RQ4** sufficiently, three classification scenarios were devised for each model:

- Sensemaking vs. The Rest: The ability to detect the presence of any sensemaking comments in the feedback text.

- Sensemaking 1 vs. Sensemaking 2: The ability to understand whether the sensemaking comments highlight specific strengths and weaknesses (Sensemaking 1) or the overall strengths and weaknesses (Sensemaking 2).

- Sensemaking and Future Impact vs. The Rest: Being able to gauge whether the feedback highlights strengths and weaknesses of the student task (Sensemaking) and simultaneously provides actionable information (Future Impact) as opposed to other rubric occurrences.

Each of these three scenarios represent a one-to-many classification problem. This means that the models had to identify one out of two outcomes. To improve model performance, we represented these two outcomes as 1 (positive class) and 0 (negative class), with 1 indicating the presence of a Sensemaking aspect and 0 indicating otherwise.

3.3.3 Handling Imbalanced Data

The data that was prepared for the second (Sensemaking 1 vs. Sensemaking 2) and third scenario (Sensemaking and Future Impact vs. The Rest) had an unbalanced target variable. To handle this disproportionate number of class occurrences in the target variable, we used resampling techniques to balance the classes. We used SMOTE (Park et al., 2023) from the imblearn-Python library to balance the data for the machine learning models. SMOTE (Park et al., 2023) balances the data by generating synthetic samples for the minority class. This gives it an advantage over other oversampling methods as the synthetic samples are not duplicates.

For the deep learning models, we used a special Keras (Ketkar, 2017) neural network library k-train-Python on top of the Tensorflow (Abadi et al., n.d.) Classifier to build penalty weights for the classes (Sensemaking 1 vs. Sensemaking 2) and (Sensemaking and Future Impact vs. The Rest). The weights penalise the majority class whenever it starts overwhelming the minority class. Overall, the class penalty method is recommended over SMOTE (Park et al., 2023) as it reduces the risk of overfitting the model (Torgo et al., 2013).

3.3.4 Training the NLP Models

We used the Scikit-Learn-Python (Li et al., 2022) library to develop and train the first three models, Logistic regression (Pavlyshenko, 2016), support vector machine and Random Forest. For implementing the XGBoost (Pavlyshenko, 2016) model, we used its dedicated package XGBoost-Python library. To deploy the DistilBERT (Sanh et al., 2020) model, we used the TensorFlow (Abadi et al., n.d.) classifier to access the pre-trained model and compile its results. The data was split with 80% for training the model and 20% for validation to test the performance of the model.

3.3.5 Evaluation Metrics

We used accuracy, precision, recall and F1-Score (Li et al., 2022) as the metrics to evaluate the classification models. They are measured from a scale of 0-1 with 1 indicating 100% classification quality. These metrics are widely accepted for classification problems (Bazazeh & Shubair, 2016).

4. Results

4.1 RQ1

We made several observations about the text with respect to its alignment with the sensemaking aspect of feedback. For the detailed analysis and code, please refer to Appendix 1.

4.1.1 Text Counts and Complexities

Overall, it was observed that the Sensemaking rubrics had the greatest diversity of words and an average readability score of 52 (See table 3) which is acceptable for class 10-12 students and very good for graduate level students (Sideris et al., 2021) (See Appendices 3.1). Additionally, the feedback comments having Sensemaking 1 and Impact 1 rubric combination had the most unique words compared to other Sensemaking combinations. The readability of the Sensemaking 2 rubric became more favourable for school-level students (Sideris et al., 2021) when it was paired with Agency 3 comments as shown in table 3. Overall, the increase in the Flesch' Reading ease score (Sideris et al., 2021) was directly proportional to the decrease in the number of unique words and sentences. This means that the readability score was higher for comments with fewer unique words and sentences.

Table 3: Text Count and Complexity Summaries

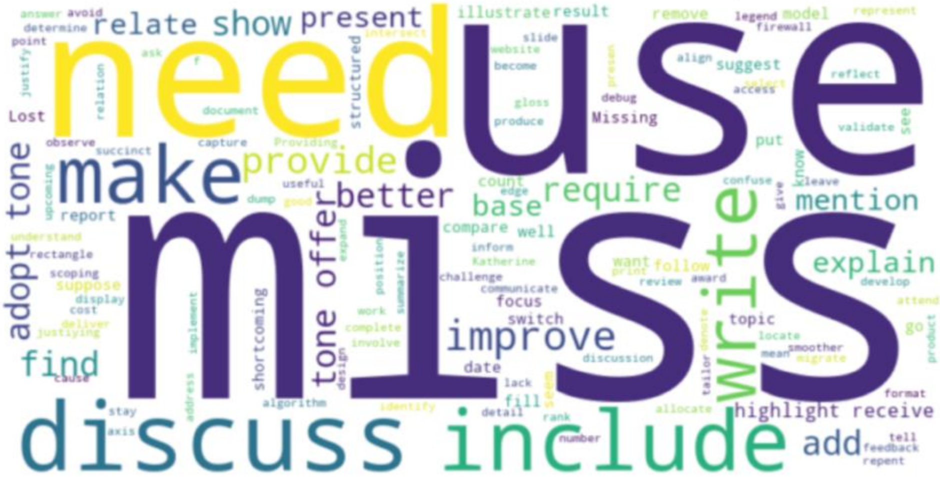
Granularity	Average Word Length	Unique Word Count	Sentence Count	Average Readability Index
Sensemaking	22.79	40024	4438	52.35
Sensemaking 1	18.60	22435	2521	52.15
Sensemaking 2	11.76	2785	413	52.46
Sensemaking 1 & Impact 1	47.40	8934	938	52.75
Sensemaking 2 & Sensemaking 1	32.56	914	77	47.25
Agency 2 & Sensemaking 1	17.47	533	52	51.91
Sensemaking 2 & Agency 3	15	8	1	92.12
Agency 2 & Sensemaking 1 & Impact 1	60.44	542	34	24.35

4.1.2 Word Combinations and Occurrences

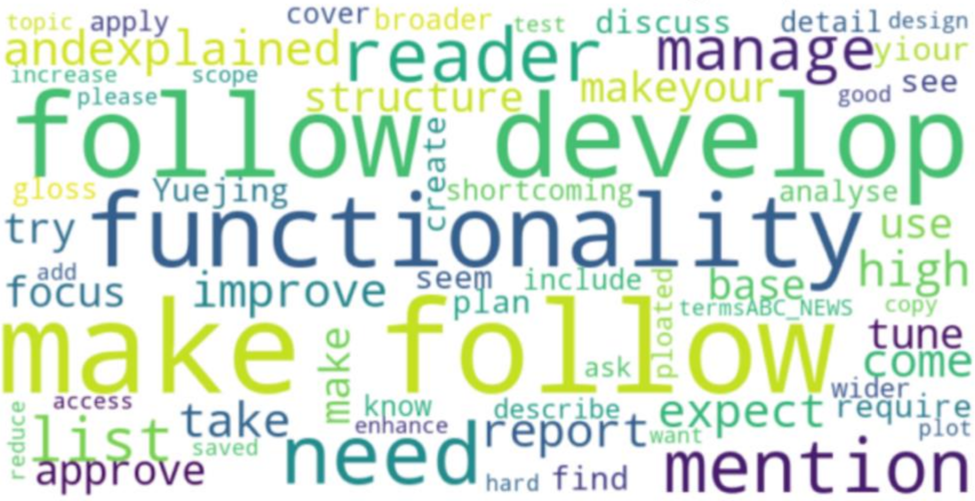
1. Common Verbs

By extracting the all the verbs for different granular levels with NLTK and representing them in the form of a word cloud, we were able to understand which verbs are being used frequently. For example, “need”, “miss” and “use” are the most common verbs for Sensemaking comments that highlight specific strengths and weaknesses.

Word Cloud of Common Verbs - Sensemaking 1



Word Cloud of Common Verbs - Sensemaking 1&Impact 1



2. Ngrams

In contrast with the common verbs, when we observe the most frequent word occurrences, we are presented with different results. For example, almost none of the verbs from the above word cloud are the most frequent words in table 4.

In fact, most of the top bigrams for Sensemaking are more affirmative (Agency 2) than actionable as shown in table 8, eg. “good work” or “well done”. Additionally, a sizable number of trigrams as shown in table 8 are phrases like “16 marks week” or “marks week 12”. Such phrases do not offer any context in terms of informing the student about strengths and weaknesses. Finally, there is a vast difference in the frequency of the most common word for the Sensemaking 1 rubric and the rubric combination of Sensemaking 1 & Impact 1, indicating fewer comments in the text that have the (Sensemaking 1 & Impact 1) comments.

Table 4: Ngrams and Word Frequencies

Granularity	Top Bigrams	Top Trigrams	Most Frequent Word	Unigram Frequency
Sensemaking	marks week, well done, good work, good job	16 marks week, 17 marks week, model evaluation inference, marks week 12	well	324
Sensemaking 1	marks week, well done, good work, good job, make sure	16 marks week, 17 marks week, model evaluation inference, marks week 12, please make sure	well	199
Sensemaking 2	marks week, well done, good work, discussion points, good job	16 marks week, 11 marks week, 18 marks week, date time location, model evaluation inference	well	83
Sensemaking 1 & Impact 1	marks week, would better, well done, good job	part notebook good, 14 marks week, model evaluation inference, lt br gt	would	87
Sensemaking 2 & Sensemaking 1	marks week, week 12, would better, 14 marks	15 marks week, marks week 12, 16 marks week, 14 marks week	well	24

4.1.3 Sensemaking Text Sentiments

For Sensemaking based comments where the Agency 2 and Agency 3 rubrics were absent, positive sentiments were more prevalent than neutral and negative sentiments. For example, almost 54% of single occurrences of Sensemaking rubrics had a positive sentiment. This is unsurprising, given the previous results where the top phrases in sensemaking comments were affirmative comments.

However, the percentage of positive sentiments was especially higher for the feedback comments that had some combination of a Sensemaking rubric with other rubrics with actionable comments (Impact 1 or Impact 2 or Impact 3) or comments that encourage students to take an active role in their feedback process (Agency 1). This indicates that there is a high percentage of positive sentiments in these other rubrics too.

Table 5: Sensemaking Sentiments

Granularity	Sentiment	Proportion (%)
Rubric Combinations with Sensemaking	Positive	68.6
	Neutral	15.2
	Negative	16.2
Single Sensemaking Rubric Occurrences	Positive	53.6
	Neutral	17.1
	Negative	29.2

4.2 RQ2

4.2.1 Grade Distribution Summary

Overall, there is a disproportion in the usage of Sensemaking comments used in the total number of feedbacks for each grade.
Firstly, Sensemaking comments were more prevalent when the student achieved grades ‘C’, ‘D’ or ‘HD’. In contrast, students who achieved the fail grade ‘N’ did not receive Sensemaking heavy comments. For example, out of 185 pieces of feedbacks that accompanied grade ‘N’, only 26 comments contained Sensemaking 2 comments which are supposed to cover the overall strengths and weaknesses of the student’s work.

Secondly, only 7 feedbacks with a fail grade had Sensemaking 1 & Impact 1 comments. In other words, out of 185 pieces of feedback only these feedbacks highlighted specific negatives of the student’s work (Sensemaking 1) that may have caused them to achieve a fail grade with actionable information (Impact 1) to improve their grades. Most of the grade ‘N’ feedbacks had solitary sensemaking comments without any impact comments. Refer to Appendices 3.2.

4.2.2 Inferential Statistics

Part 2 of **RQ2** involved understanding the dependency between the student score and different granularity levels. Overall, the Chi-Squared (Varvara et al., 2021) test showed sufficient evidence (i.e. p_value < 0.01) to indicate that there is a dependency between the labels and the student grade as shown in table 6.

Additionally, the comments with multi-rubric Sensemaking comments also showed a significant dependency on the grade (i.e. p-value < 0.01). This shows that there is a strong dependency between multi-rubric Sensemaking comments alone with the student grade.

Table 6: Inferential Statistics

Granularity	Chi-Squared Test Results	
Component Level	Statistic	78.23
	p-value	< 0.01
	Degrees of Freedom	8
Single Rubric Level	Statistic	206.39
	p-value	< 0.01
	Degrees of Freedom	28
Single Rubric Level - Sensemaking	Statistic	20.92
	p-value	< 0.01
	Degrees of Freedom	4
Multi-Rubric Level	Statistic	360.97
	p-value	< 0.01
	Degrees of Freedom	168
Multi-Rubric Level - Sensemaking	Statistic	138.22
	p-value	< 0.01
	Degrees of Freedom	88

4.3 RQ3

Overall, we discovered that Sensemaking comments satisfied all six levels of Bloom's taxonomy as shown in table 7, indicating an extensivity in the variety of action verbs being used. This was extended to the association between Sensemaking 1 and Impact 1 but not to Agency 1 or Agency 2 where less than 6 levels were aligned.

Table 7: Bloom's Taxonomy-Sensemaking

Granularity	Bloom's Taxonomy Levels (Pickard, 2007)	Number of Levels
Sensemaking	Remembering, Evaluating, Creating, Applying, Understanding, Analyzing	**6**
Sensemaking 1	Evaluating, Creating, Applying, Remembering, Understanding, Analyzing	**6**
Sensemaking 2	Applying, Creating, Evaluating, Remembering, Understanding, Analyzing	**6**
Sensemaking 1 & Impact 1	Remembering, Applying, Creating, Analyzing, Understanding, Evaluating	**6**
Sensemaking 1 & Agency 1	Remembering, Understanding, Evaluating, Creating, Applying	**5**
Sensemaking 2 & Sensemaking 1 & Impact 1	Applying, Remembering, Understanding, Analyzing, Evaluating	**5**
Agency 2 & Sensemaking 2	Evaluating, Creating, Applying	**3**

4.4 RQ4

A considerably favourable result was achieved in the identification of the Sensemaking (Ryan et al., 2021) aspect. The deep learning models with DistilBERT (Sanh et al., 2020) were the best performing models.

4.4.1 Machine Learning Classification

Overall, the TF-IDF (Nasim et al., 2017) feature extraction method improved the model performance by a considerable margin compared to LIWC (Boyd et al., 2022), performing exceptionally on multiple parameters such as Accuracy, and F1-Score. The result from the best models is summarised in table 8.

For the Sensemaking vs. the Rest scenario, a maximum accuracy of 87% was achieved with the TF-IDF-Support Vector Machine model. The recall of 87% in comparison with the accuracy indicates that the target classes were also balanced. In the case of the Sensemaking 1 vs. Sensemaking 2 scenario, the TF-IDF-Random Forest model had the best performance with a 94% accuracy. Given that the model had an acceptable recall of 83%, it shows that the oversampling method SMOTE did not cause severe overfitting. Finally, for identifying

Sensemaking & Future Impact vs. the Rest, a 95% accuracy was achieved with LIWC-XGBoost. Additional diagrams of the ROC curves can be found in appendices section 2.1.

Table 8: Sensemaking Classification Result for Machine Learning Models

Scenario	Feature Extraction	Model	Accuracy	Precision	Recall	F1-Score
Sensemaking vs The Rest	TF-IDF	Logistic Regression	0.85	0.85	0.85	0.85
	TF-IDF	Support Vector Machine	0.87	0.87	0.87	0.86
	TF-IDF	Random Forest	0.86	0.86	0.86	0.86
	LIWC	XGBoost	0.86	0.86	0.85	0.86
Sensemaking 1 vs. Sensemaking 2	TF-IDF	Logistic Regression	0.93	0.84	0.83	0.83
	TF-IDF	Support Vector Machine	0.94	0.89	0.82	0.85
	TF-IDF	Random Forest	0.94	0.89	0.83	0.86
	TF-IDF	XGBoost	0.93	0.88	0.81	0.84
Sensemaking & Future Impact vs. The Rest	TF-IDF	Logistic Regression	0.89	0.68	0.79	0.71
	TF-IDF	Support Vector Machine	0.94	0.81	0.71	0.75
	LIWC	Random Forest	0.94	0.81	0.80	0.81
	LIWC	XGBoost	0.95	0.83	0.80	0.81

4.4.2 LIWC Feature Importance

Overall, the LIWC (Boyd et al., 2022) extracted features were not sufficient to train the classification models as well as the TF-IDF (Nasim et al., 2017) extracted features. However, the LIWC-XGBoost model gave the best results for model types 1 and 3. Therefore, we have presented the top ten LIWC (Boyd et al., 2022) features in table 9 for these model types with their linguistic categories. This has large implications for future studies (See section 6).

Table 9: LIWC-XGBoost Feature Importance

Scenario	Category (Boyd et al., 2022)	Sub-Category (Boyd et al., 2022)	Feature (Boyd et al., 2022)
Sensemaking vs The Rest	Linguistic Dimensions	Total Function Words	verb
	Linguistic Dimensions	Total Function Words	Auxverb
	Expanded Dictionary	Time Orientation	focuspresent
	Summary Variables	Word Count	Dic
	Psychological Processes	Cognition	discrep
	Expanded Dictionary	Motives	allure
	Psychological Processes	Social Processes	Social
	Time Orientation	Time Orientation	focuspast
	Linguistic dimensions	Total Function Words	prep
	Summary variables	Word Count	BigWords
Sensemaking & Future Impact vs. The Rest	Summary variables	Word Count	WC
	Summary variables	Word Count	WPS
	Psychological Processes	Cognition	discrep
	Expanded Dictionary	All Punctuation	Comma
	Expanded Dictionary	Time orientation	focuspresent
	Summary variables	Word Count	Analytic
	Summary variables	Word Count	Dic
	Summary variables	Word Count	Authentic
	Summary variables	Word Count	Tone
	Expanded Dictionary	Perception	motion

4.4.3 Deep Learning Classification

Overall, the DistilBERT (Sanh et al., 2020) model types outperformed the machine learning models with a 90% average accuracy as shown in table 10. Additionally, the model for identifying Sensemaking & Impact comments gave a high precision and recall (Li et al., 2022) of 95% and 96% respectively indicating that a sizable quantity of the text having Sensemaking and Impact comments was correctly identified.

However, the deep learning model for the second case (Sensemaking 1 vs. Sensemaking 2) produced a slightly lower precision and recall at 87% and 81% respectively compared to the TF-IDF-Random Forest model (See table 8).

Table 10: DistiBERT (Sanh et al., 2020) Results

Scenario	Accuracy	Precision	Recall	F1-Score
Sensemaking vs. The Rest	0.89	0.91	0.89	0.90
Sensemaking 1 vs. Sensemaking 2	0.93	0.87	0.81	0.83
Sensemaking & Future Impact vs. The Rest	0.93	0.98	0.95	0.96

5. Discussion

Our findings have large implications for learner-centred feedback. Sufficient evidence was found to support the claim (Ryan et al., 2021) that Sensemaking can facilitate Future Impact (Ryan et al., 2021), given that (Sensemaking 1 & Impact 1) was the most dominant rubric combination in the data in terms of unique word count and sentence count. However, individual occurrences of the Sensemaking 1 rubric were exponentially higher than (Sensemaking 1 & Impact 1), indicating that most feedbacks in practice have a lot of comments highlighting specific strengths and weaknesses (Sensemaking 1) but they fail to provide actionable information (Impact 1) to improve upon those strengths and weaknesses.

Additionally, we found evidence showing the use of affirmative phrases in Sensemaking comments. For example, a few of the commonly used phrases found in Sensemaking comments such as “well done” or “good work” are more frequent in Sensemaking comments than informative or actionable verb phrases drawing attention to strengths and weaknesses such as “please make sure” or “would better”. This finding was solidified further by sentiment analysis which discovered a large percentage of positive sentiments in the text. While positivity helps the student engage with feedback, teachers can be guided also include more actionable information (Dohrer, 1991; Esterhazy and Damşa, 2019; Zhang and Zheng 2018; Ryan et al., 2022) with comments that highlight strengths and weaknesses (Ryan et al., 2021). Based on the top word occurrences, we have established a tangible way to create an automated blueprint to help teachers avoid overusing certain words and phrases.

The feedback also suffers from lack of complexity when the student achieves a ‘fail’ grade. For example, only 7 grade ‘N’ feedbacks were found with (Sensemaking & Future impact) comments. This finding was supported by the Chi-squared (Varvara et al., 2021) test which showed that the student’s performance (grade) is determining the complexity of the feedback text. Therefore, educators can be recommended to give more detailed and actionable comments for failing students.

Our novel method for aligning Sensemaking (Ryan et al., 2021) with Bloom's Taxonomy (Krathwohl, 2002) is significant in terms of verb associations. By measuring the alignment between the action verbs in the Sensemaking comments and Bloom's Taxonomy levels (Pickard, 2007), specific cognitive levels can be assigned to the text and function as a ranking for the educator's cognitive ability.

Additionally, research that analyses Bloom's Taxonomy can repurpose our technique to label different types of educational texts and use it to build complex classification models. This will save the time it took for some researchers such as Lin et al. (Lin et al., 2022) to manually label the data. Moreover, our method can be combined with Lin et al.'s (Lin et al., 2022) study to use Bloom's Taxonomy as a barometer for studying the alignment of Sensemaking comments with learning objectives.

With DistilBERT (Sanh et al., 2020), we were able to achieve a high level of classification quality. This bodes well for future iterations of this model which can be trained against larger datasets. Moreover, this finding complements previous studies by showing that deep learning algorithms comparatively perform better in classifying education texts compared to traditional machine learning models (Sha et al., 2021; Li et al., 2022). Finally, a key aspect of these models is their reusability. For example, a comprehensive DistilBERT-based feedback model can be saved and shared instantly with interested education institutions. They can easily leverage their automated nature to identify the learner-centred (Ryan et al., 2021) components in their feedback.

6. Limitations and Future Work

We made significant progress in providing a focused approach to analysing and detecting learner-centred feedback through sensemaking (Ryan et al., 2021). However, due to the scope of our study, a few aspects could not be studied. This study can be expanded in future research projects with the following additions:

- **Data Diversity**: The feedback data being used in this study was limited to the units of the Master of Data Science course. Future studies may extend this to different disciplines and note how the Sensemaking aspect varies between them.
- **Data Size**: Due to the enormous task of labelling a randomized sample of the data, only 1000 feedback pieces were used. We have an untapped master dataset with more than 50000 unique pieces. A larger team can be utilised for the labelling of many samples in future research.
- **Class Imbalance**: There is a risk of overfitting the data with class balancing techniques. Fortunately, we achieved a good accuracy and recall with our classification models. However, future studies may benefit from a large dataset by avoiding the use of techniques like SMOTE and class penalty.
- **Focus on Future Impact and Support Agency** (Ryan et al., 2021): This project mainly focused on the sensemaking (Ryan et al., 2021) aspect of learner-centred feedback. However, the methodology implemented for this study can be expanded to include the other two components of learner-centred feedback in upcoming studies.

- **Reusing LIWC features:** The LIWC (Boyd et al., 2022) extracted features improved the performance of the XGBoost (Pavlyshenko, 2016) model for model scenarios 1 and 3. Considering that the use of LIWC (Boyd et al., 2022) saw promising results in similar studies (Lin et al., 2023) and covers a sizable corpus of features, future studies in this research can reuse it with a fresh set of feedback samples. The top ten features can also be analysed further in the future to understand whether they are positively or negatively correlated (Lin et al., 2023) with the data.

7. Conclusion

This study has taken an important step to bridge the gap between the conceptual understanding of the components of learner-centred feedback and their potential applicability in current feedback practices. Through our findings, we were able to confirm aspects of the text such as its readability that aligned with the Sensemaking aspect. We also found areas for improvement such as the inclusion of more actionable information with sensemaking comments and detailed feedback for students who have a fail grade. We were also able to showcase the ability of NLP deep learning models to take a feedback text and classify it as a Sensemaking comment with a 90-95% certainty. These findings have large implications for future studies on feedback in terms of automating the classification of learner-centred feedback components within feedback practice.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., … Zheng, X. (n.d.). *TensorFlow: A system for large- scale machine learning*. https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf

Bazazeh, D., & Shubair, R. (2016). Comparative study of machine learning algorithms for breast cancer detection and diagnosis. *2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA)*, 1–4. https://doi.org/10.1109/ICEDSA.2016.7818560

Botelho, A., Baral, S., Erickson, J. A., Benachamardi, P., & Heffernan, N. T. (2023). Leveraging natural language processing to support automated assessment and feedback for student open responses in mathematics. *Journal of Computer Assisted Learning*, *39*(3), 823–840. mathematics. *Journal of Computer Assisted Learning*, *39*(3), 823–840. https://doi.org/10.1111/jcal.12793

Boud, D., & Dawson, P. (2023). What feedback literate teachers do: An empirically-derived competency framework. *Assessment and Evaluation in Higher Education*, *48*(2), 158–171. Scopus. https://doi.org/10.1080/02602938.2021.1910928

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment and Evaluation in Higher Education*, *38*(6), 698–712. Scopus. https://doi.org/10.1080/02602938.2012.691462

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of LIWC-22. Austin, TX: University of Texas at Austin. https://www.liwc.app

Brooks, C., Burton, R., van der Kleij, F., Carroll, A., & Hattie, J. (2021). Towards student-centred feedback practices: Evaluating the impact of a professional learning intervention in primary schools. *Assessment in Education: Principles, Policy & Practice, 28*(5–6), 633–656. https://doi.org/10.1080/0969594X.2021.1976108

Chang, W.-C., & Chung, M.-S. (2009). *Automatic applying Bloom's taxonomy to classify and analysis the cognition level of english question items*. 727–733. Scopus. https://doi.org/10.1109/JCPC.2009.5420087

Chiche, A., & Yitagesu, B. (2022). Part of speech tagging: A systematic review of deep learning and machine learning approaches. *Journal of Big Data*, *9*(1), 10. https://doi.org/10.1186/s40537-022-00561-y

Das, S., Das Mandal, S. K., & Basu, A. (2022). Classification of Action Verbs of Bloom's Taxonomy Cognitive Domain: An Empirical Study. *Journal of Education*, *202*(4), 554–566. Scopus. https://doi.org/10.1177/00220574211002199

Dawson, P., Henderson, M., Mahoney, P., Phillips, M., Ryan, T., Boud, D., & Molloy, E. (2019). What makes for effective feedback: Staff and student perspectives. *Assessment & Evaluation in Higher Education*, *44*(1), 25–36. https://doi.org/10.1080/02602938.2018.1467877

Dawson, P., Henderson, M., Ryan, T., Mahoney, P., Boud, D., Phillips, M., & Molloy, E. (2018). Technology and feedback design. In M. J. Spector, B. B. Lockee, & M. D. Childress (Eds.), Learning, Design, and Technology (pp. 1–45). Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-17727-4 124-1

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. http://arxiv.org/abs/1810.04805

Dohrer, G. (1991). Do Teachers' Comments on Students' Papers Help? *College Teaching*, *39*(2), 48–54. https://doi.org/10.1080/87567555.1991.9925485

Evans, C. (2013). Making sense of assessment feedback in higher education. Review of Educational Research, 83(1), 70–120. https://doi.org/10.3102/0034654312474350

Esterhazy, R., & Damşa, C. (2019). Unpacking the feedback process: An analysis of undergraduate students' interactional meaning-making of feedback comments. *Studies in Higher Education*, *44*(2), 260–274. https://doi.org/10.1080/03075079.2017.1359249

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221–233. https://doi.org/10.1037/h0057532

Hattie, J. (2008). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. Routledge. https://doi.org/10.4324/9780203887332

Hattie, J., Gan, M., & Brooks, C. (2016). Instruction based on feedback. *Handbook of Research on Learning and Instruction (2nd Ed.).* New York, NY: Routledge

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research, 77*(1), 81-112. https://doi.org/10.3102/003465430298487

Henderson, M., Ajjawi, R., Boud, D., & Molloy, E. (2019). Identifying Feedback That Has Impact. In *The Impact of Feedback in Higher Education: Improving Assessment Outcomes for Learners* (pp. 15–34). Springer International Publishing. https://doi.org/10.1007/978-3-030-25112-3_2

Henderson, M., Phillips, M., Ryan, T., Boud, D., Dawson, P., Molloy, E., & Mahoney, P. (2019). Conditions that enable effective feedback. *Higher Education Research & Development*, *38*(7), 1401–1416. https://doi.org/10.1080/07294360.2019.1657807

Iraj, H., Fudge, A., Khan, H., Faulkner, M., & Pardo, A. (2021). Narrowing the Feedback Gap: Examining Student Engagement with Personalized and Actionable Feedback Messages. *Journal of Learning Analytics*, *8*(3), 101–116. Scopus. https://doi.org/10.18608/jla.2021.7184

Karttunen, L., Chanod, J.-P., Grefenstette, G., & Schille, A. (1996). Regular expressions for language engineering. *Natural Language Engineering*, *2*(4), 305–328. https://doi.org/10.1017/S1351324997001563

Katz, A., Norris, M., Alsharif, A. M., Klopfer, M. D., Knight, D. B., & Grohs, J. R. (2021). *Using Natural Language Processing to Facilitate Student Feedback Analysis*. ASEE Annual Conference and Exposition, Conference Proceedings. Scopus. https://peer.asee.org/using-natural-language-processing-to-facilitate-student-feedback-analysis

Ketkar, N. (2017). Introduction to Keras. In N. Ketkar (Ed.), *Deep Learning with Python: A Hands-on Introduction* (pp. 97–111). Apress. https://doi.org/10.1007/978-1-4842-2766-4_7

Koşar, G. (2021). The Progress a Pre-Service English Language Teacher Made in Her Feedback Giving Practices in Distance Teaching Practicum. *JET (Journal of English Teaching), 7*(3), 366–381. https://doi.org/10.33541/jet.v7i3.3145

Krathwohl, D. R. (2002). A Revision of Bloom’s Taxonomy: An Overview. *Theory Into Practice*, *41*(4), 212–218. https://www.jstor.org/stable/1477405

Lee, I., Mak, P., & Yuan, R. E. (2019). Assessment as learning in primary writing classrooms: An exploratory study. Studies in Educational Evaluation, 62, 72–81. https://doi.org/10.1016/j.stue duc.2019.04.012

Li, Y., Rakovic, M., Poh, B. X., Gaševic, D., & Chen, G. (2022). Automatic Classification of Learning Objectives Based on Bloom’s Taxonomy. In *International Educational Data Mining Society*. International Educational Data Mining Society. https://eric.ed.gov/?id=ED624058

Lim, L.-A., Dawson, S., Gašević, D., Joksimović, S., Pardo, A., Fudge, A., & Gentili, S. (2021). Students’ perceptions of, and emotional responses to, personalised learning analytics-based feedback: An exploratory study of four courses. *Assessment & Evaluation in Higher Education*, *46*(3), 339–359. https://doi.org/10.1080/02602938.2020.1782831

Lin, J., Dai, W., Lim, L.-A., Tsai, Y.-S., Mello, R. F., Khosravi, H., Gasevic, D., & Chen, G. (2023). *Learner- centred Analytics of Feedback Content in Higher Education*. 100–110. Scopus. https://doi.org/10.1145/3576050.3576064

Lipsch-Wijnen, I., & Dirkx, K. (2022). A case study of the use of the Hattie and Timperley feedback model on written feedback in thesis examination in higher education. *Cogent Education*, *9*(1). Scopus. https://doi.org/10.1080/2331186X.2022.2082089

Mandouit, L., & Hattie, J. (2023). Revisiting "The Power of Feedback" from the perspective of the learner. *Learning and Instruction*, *84*. Scopus. https://doi.org/10.1016/j.learninstruc.2022.101718

Nasim, Z., Rajput, Q., & Haider, S. (2017). Sentiment analysis of student feedback using machine learning and lexicon based approaches. *2017 International Conference on Research and Innovation in Information Systems (ICRIIS)*, 1–6. https://doi.org/10.1109/ICRIIS.2017.8002475

Mishra, P., Singh, U., Pandey, C. M., Mishra, P., & Pandey, G. (2019). Application of Student's t- test, Analysis of Variance, and Covariance. *Annals of Cardiac Anaesthesia*, *22*(4), 407–411. https://doi.org/10.4103/aca.ACA_94_19

Osadi, A., Fernando, N., & Welgama, V. (2017). Ensemble Classifier based Approach for Classification of Examination Questions into Bloom's Taxonomy Cognitive Levels. *International Journal of Computer Applications*, *162*, 975–8887. https://doi.org/10.5120/ijca2017913328

Park, J., Kwon, S., & Jeong, S.-P. (2023). A study on improving turnover intention forecasting by solving imbalanced data problems: Focusing on SMOTE and generative adversarial networks. *Journal of Big Data*, *10*(1), 36. https://doi.org/10.1186/s40537-023-00715-6

Pavlyshenko, B. (2016). Machine learning, linear and Bayesian models for logistic regression in failure detection problems. *2016 IEEE International Conference on Big Data (Big Data)*, 2046–2050. https://doi.org/10.1109/BigData.2016.7840828

Pickard, M. J. (2007). The new Bloom's taxonomy: An overview for family and consumer sciences. *Journal of Family and Consumer Sciences Education*, *25*(1). https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=d24bb1ad0754e7d59eb7e3e28614bba2b6874180

R, N., S, P. M., Harithas, Pramath. P., & Hegde, V. (2022). Sentimental Analysis on Student Feedback using NLP & POS Tagging. *2022 International Conference on Edge Computing and Applications (ICECAA)*, 309–313. https://doi.org/10.1109/ICECAA55415.2022.9936569

Ryan, T., Henderson, M., Ryan, K., & Kennedy, G. (2021). Designing learner-centred text-based feedback: A rapid review and qualitative synthesis. *Assessment & Evaluation in Higher Education*, *46*(6), 894–912. https://doi.org/10.1080/02602938.2020.1828819

Ryan, T., Henderson, M., Ryan, K., & Kennedy, G. (2021). Identifying the components of effective learner-centred feedback information. *Teaching in Higher Education*, *0*(0), 1–18. https://doi.org/10.1080/13562517.2021.1913723

Ryan, T., Henderson, M., Ryan, K., & Kennedy, G. (2022). Feedback in higher education: Aligning academic intent and student sensemaking. *Teaching in Higher Education*. Scopus. https://doi.org/10.1080/13562517.2022.2029394

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter* (arXiv:1910.01108). arXiv. https://doi.org/10.48550/arXiv.1910.01108

Sha, L., Rakovic, M., Li, Y., Whitelock-Wainwright, A., Carroll, D., Gaševic, D., & Chen, G. (2021). Which Hammer Should I Use? A Systematic Evaluation of Approaches for Classifying Educational Forum Posts. In *International Educational Data Mining Society*. International Educational Data Mining Society. https://eric.ed.gov/?id=ED615664

Shaik, T., Tao, X., Li, Y., Dann, C., McDonald, J., Redmond, P., & Galligan, L. (2022). A Review of the Trends and Challenges in Adopting Natural Language Processing Methods for Education Feedback Analysis. *IEEE Access*, *10*, 56720–56739. Scopus. https://doi.org/10.1109/ACCESS.2022.3177752

Sideris, G. A., Vyllioti, A.-T., Dima, D., Chill, M., & Njuguna, N. (2021). The Value of Web- Based Patient Education Materials on Transarterial Chemoembolization: Systematic Review. *JMIR Cancer*, *7*(2), e25357. https://doi.org/10.2196/25357

Sulis, E., Humphreys, L., Vernero, F., Amantea, I. A., Audrito, D., & Di Caro, L. (2022). Exploiting co-occurrence networks for classification of implicit inter-relationships in legal texts. *Information Systems*, *106*, 101821. https://doi.org/10.1016/j.is.2021.101821

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Timperley, H. S., & Parr, J. M. (2009). What is this lesson about? Instructional processes and student understandings in writing classrooms. *The Curriculum Journal*, *20*(1), 43–60. https://doi.org/10.1080/09585170902763999

Torgo, L., Ribeiro, R. P., Pfahringer, B., & Branco, P. (2013). SMOTE for Regression. In L. Correia, L. P. Reis, & J. Cascalho (Eds.), *Progress in Artificial Intelligence* (pp. 378–389). Springer. https://doi.org/10.1007/978-3-642-40669-0_33

Tsai, Y.-S., Mello, R. F., Jovanović, J., & Gašević, D. (2021). *Student appreciation of data-driven feedback: A pilot study on OnTask*. 511–517. Scopus. https://doi.org/10.1145/3448139.3448212

Varvara, G., Bernardi, S., Bianchi, S., Sinjari, B., & Piattelli, M. (2021). Dental Education Challenges during the COVID-19 Pandemic Period in Italy: Undergraduate Student Feedback, Future Perspectives, and the Needs of Teaching Strategies for Professional Development. *Healthcare*, *9*(4), Article 4. https://doi.org/10.3390/healthcare9040454

Wiliam, D. (2017). Embedded formative assessment (2nd ed.). Solution Tree Press.

Winstone, N., Boud, D., Dawson, P., & Heron, M. (2022). From feedback-as-information to feedback-as- process: A linguistic analysis of the feedback literature. *Assessment and*

Evaluation in Higher Education, *47*(2), 213–230. Scopus. https://doi.org/10.1080/02602938.2021.1902467

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., … Rush, A. M. (2020). *HuggingFace's Transformers: State-of-the- art Natural Language Processing* (arXiv:1910.03771). arXiv. https://doi.org/10.48550/arXiv.1910.03771

Yang, J., Zhang, Y., Li, L., & Li, X. (2018). *YEDDA: A Lightweight Collaborative Text Span Annotation Tool* (arXiv:1711.03759). arXiv. https://doi.org/10.48550/arXiv.1711.03759

Yang, M., & Carless, D. (2013). The feedback triangle and the enhancement of dialogic feedback processes. *Teaching in Higher Education*, *18*(3), 285–297. Scopus. https://doi.org/10.1080/13562517.2012.719154

Zhang, L., & Zheng, Y. (2018). Feedback as an assessment for learning tool: How useful can it be? *Assessment & Evaluation in Higher Education*, *43*(7), 1120–1132. https://doi.org/10.1080/02602938.2018.1434481

Zhang, X., Mao, R., & Cambria, E. (2023). A survey on syntactic processing techniques. *Artificial Intelligence Review*, *56*(6), 5645–5728. https://doi.org/10.1007/s10462-022-10300-7

Appendices

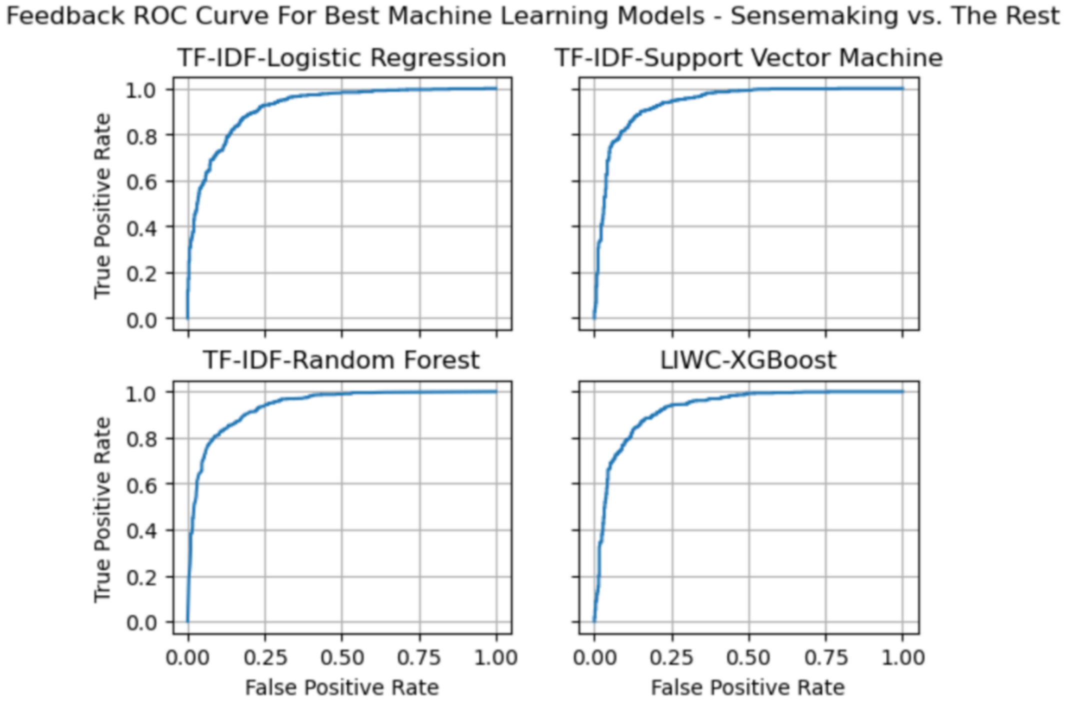
1. Code:

The code that was used to answer the research questions also contains secondary findings which were not included in this report. Please note, the code is in a '.ipynb' format. You will need a compatible tool such as Jupyter notebook to view them. It can be accessed with this link: https://github.com/Siddharth1989/LearnerCentricFeedbackEnhancement.git. Due to Monash University's data confidentially issues and on the recommendation of my supervisors, the raw data and the pre-processing code cannot be shared in the above repository. In case of any questions regarding the raw data or code, please reach out to me on sgup0021@student.monash.edu

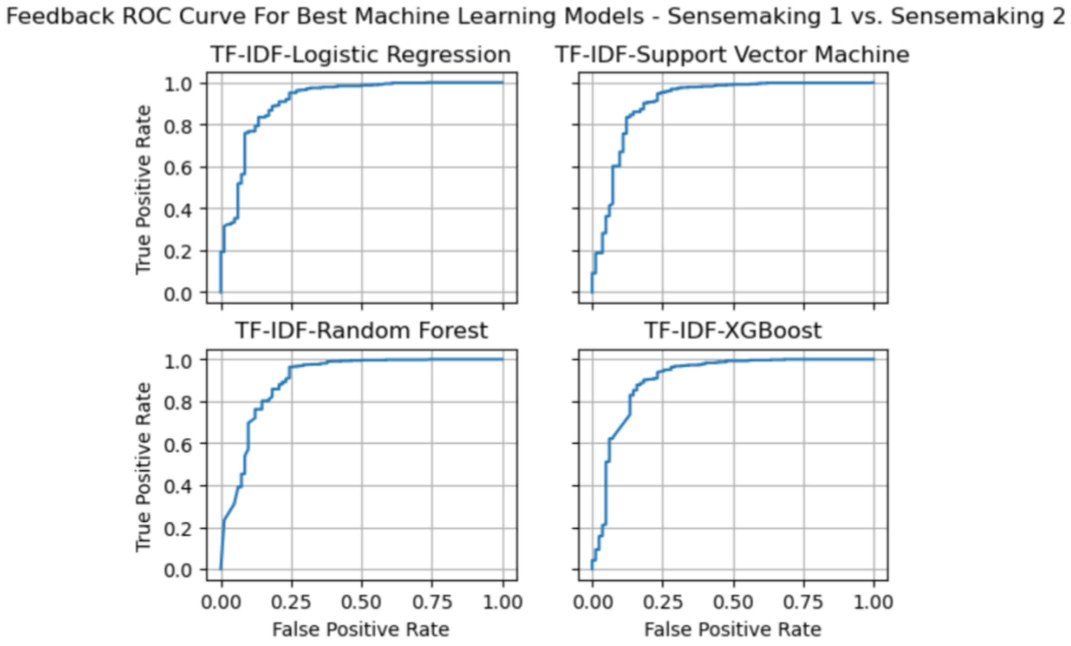
2. Figures

2.1 ROC Curves

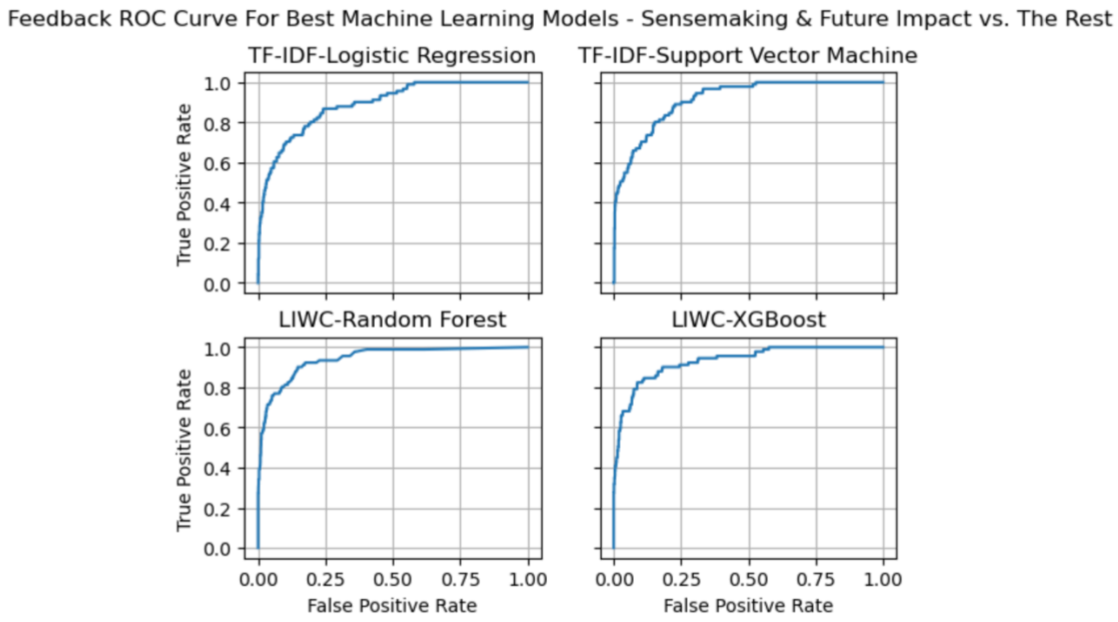
2.1.1 Sensemaking vs. The Rest



2.1.2 Sensemaking 1 vs. Sensemaking 2



2.1.3 Sensemaking & Future Impact vs. The Rest



3. Tables

3.1 Flesch's Reading Ease Chart (Sideris et al., 2021)

Score	Readability Level
0-30	Graduate Level
30-50	College Level
50-60	Class 10-12
60-70	Class 8-9
70-80	Class 7
80-90	Class 6
90-100	Class 5

3.2 Grade Distribution Summary

Granularity	C	D	HD	N	Total
Sensemaking	524	747	654	95	2020
Sensemaking 1	351	523	462	56	1392
Sensemaking 2	49	73	85	26	233
Sensemaking 1 & Impact 1	99	108	73	7	287
Sensemaking 2 & Sensemaking 1	11	16	14	1	42
Grade Count	1034	1467	1288	185	3974