Local Topic Mining for Reflective Medical Writing

Amy L. Olex, M.S.[1], Deborah DiazGranados, Ph.D.[1], Bridget T. McInnes, Ph.D.[1], and Stephanie Goldberg, M.D.[2]
[1]Virginia Commonwealth University, Richmond, VA, USA; [2]Virginia Commonwealth University Health System, Richmond, VA, USA

Abstract

Reflective writing is used by medical educators to identify challenges and promote inter-professional skills. These non-medical skills are central to leadership and career development, and are clinically relevant and vital to a trainees success as a practicing physician. However, identification of actionable feedback from reflective writings can be challenging. In this work, we utilize a Natural Language Processing pipeline that incorporates a seeded Term Frequency-Inverse Document Frequency matrix along with sentence-level summarization, sentiment analysis, and clustering to organize sentences into groups, which can aid educators in assessing common challenges experienced by Acting Interns. Automated analysis of reflective writing is difficult due to its subjective nature; however, our method is able to identify known and new challenges such as issues accessing the electronic health system and adjusting to specialty differences. Medical educators can utilize these topics to identify areas needing attention in the medical curriculum and help students through this transitional time.

1 Introduction

Reflective writing helps one to understand their own learning process by providing a medium with which to reflect on, assimilate, and integrate life experiences. Medical schools utilize reflective writing to help students develop non-medical skills, such as empathy, collaboration, communication, and professionalism[1,2], that are central to leadership and career development, and are clinically relevant and vital to a trainees success as a practicing physician.

The transition from a medical student to a practicing intern is a stressful process with many professional and personal challenges to overcome[3–6]. The Acting Internship (AI) program at Virginia Commonwealth University Health System aids in this transition by allowing students the graduated responsibility and autonomy to care for patients, and an opportunity to simulate the role of an intern. Medical educators are continually looking for ways to improve the AI experience to ease student transition to a practicing intern and to give them the most relevant experience by adapting the program to changing technology and clinical advancements. Reflective writing is utilized to help students assimilate new experiences encountered while in the AI role, and to aid educators in identifying challenges experienced by students while in the program. Hundreds of writings are collected each year, however, manual analysis by medical educators (who also have clinical responsibilities) is time consuming and subject to educator bias. In this paper, we introduce a novel Natural Language Processing (NLP) pipeline that is used to automatically organize reflections by common themes to aid educators in quickly identifying actionable items.

Analysis of reflective writing is challenging due to its personalized and unstructured nature[7]. Previous work on utilizing NLP to analyze reflective writing is limited and includes the use of topic modeling methods and machine learning techniques. Latent Dirichlet Allocation (LDA)[8], a popular topic modeling technique, was used by Chen et. al.[9] to identify the progression of themes in pre-service teachers reflective writing on a weekly basis, and Gibsion and Kitto[10] utilized LDA to suggest possible classification features as part of their anomaly recontextualization process when analyzing reflective texts from first year Bachelor IT students. Vrana et. al.[11] utilized Latent Semantic Analysis (LSA)[12], a predecessor of LDA, at a sentence level to analyze the coherence of expressive patient narratives about traumatic events. Finally, a supervised machine learning approach was implemented by Poon et. al.[13] to classify student's reflective writing into pre-defined topic groups, however, their approach requires topics to be defined ahead of time. To our knowledge, applying NLP to medical student reflective writing has not been reported on before.

The goal of this work is to identify both known and new challenges experienced by AI students utilizing their reflective writings. A fully supervised approach would not be appropriate as we would need to specify the challenges we are looking for ahead of time and would not discover anything new. However, implementing an unsupervised exploratory approach, while useful in aiding humans in navigating large corpora to identify general trends and themes[14,15], assumes we do not know the underlying structure of the data, which is not completely true as we do have some information on challenges many AIs experience. Therefore, this study utilizes known challenges to seed a Term Frequency-Inverse Document Frequency (TF-IDF) matrix[16] (seed topics) along with the corpus being analyzed. This seeding serves to give more weight to known topics, but also allows new topics to emerge. Our NLP pipeline utilizes the seeded TF-IDF along with sentence-level summarization, sentiment analysis, and clustering to identify groups of sentences expressing a common theme. LDA is then performed on sentence clusters to identify key words associated with each.

The following sections first describe the corpus utilized in this study, including the steps taken to clean the data for NLP processing. Next, a description of the seed topics is provided followed by the annotation of the Development and Test Corpora. The last two sections of Methods include a detailed description of the NLP pipeline implementation along with how its performance is evaluated. Following the Methods section, the results from running our pipeline on the Development and Test Corpora is described, and a discussion of these results and the annotation process is provided in the Discussion section. Finally, limitations, future work and conclusions are provided.

2 Methods

2.1 Acting Intern Corpus

The corpus used in this study is composed of reflective writings from 651 medical students in their fourth-year Acting Internship. Reflective writings were collected from blog posts and email threads, where the instructor started each thread with a prompt. Responses to the prompt *"What challenges have you faced thus far in the Acting Internship role? (for example, incorporating into the team, understanding your role, achieving goals and objectives, transitioning to the AI role, etc)"*, referred to as *challenge responses*, are used in this study to identify common challenges encountered by medical interns. To clean the corpus, we removed the initial post for each instructor-initiated thread, all student-initiated threads, direct quotes of the prompt in intern responses, and direct quotes of previous posts so that a response is not double counted. There were also a few instances where the student posted their response to an incorrect thread, such as responding to the wrong prompt, or posting to an old thread. In all identified cases, the student re-posted to the correct thread and indicated their mistake in the text. These incorrect postings were also removed from the corpus. The final corpus of challenge responses contains a total of 665 reflective writings from August 2016 to July 2018.

The corpus of responses is organized in blocks where each month is a new set of students (there are a few exceptions where the same student participated in multiple AIs). Two prompts are provided approximately 2 weeks apart with the challenge prompt being given in the first week of each AI block. A Development and Test Corpus were extracted from the full corpus to train and evaluate the NLP pipeline. The Development Corpus includes 14 manually selected challenge responses from an email thread with a total of 172 sentences. These responses were selected specifically because they express known challenges students experience, including feeling overwhelmed, having/lack of confidence, experiencing technical system issues, and having a supportive environment. The Test Corpus is composed of a single, blindly chosen block of challenge responses from March 2017 that is about the same size as the Development Corpus, and includes 22 challenge responses with a total of 155 sentences.

2.2 Seed Topics

Seed topics emphasize known challenges medical students encounter during their Acting Internship (AI), and were manually chosen by domain experts (see next section) using the Development Corpus of challenge responses (Table 1). A document for each seed topic was created by extracting all relevant sentences from the Development Corpus and incorporating them into a single document. This concentrates the terms associated with each topic within a single file, which influences the scores these terms receive in the Term Frequency-Inverse Document Frequency (TF-IDF) matrix. We anticipate that at least one of these seed topics will appear in any subset of challenge responses chosen from the entire corpus, which should result in a cluster of sentences containing most of the associated seed topic sentences along with additional sentences from the input challenge responses. Seed topics are merged with the input corpus prior to any preprocessing and TF-IDF creation, and they are removed prior to calculating F1, Precision, and Recall measures. Seed topics are meant to give more weight to these known topics during construction of sentence representations and to guide the researcher in manual analysis of the resulting clusters; however, this method does not prevent identification of new topics. To our knowledge, seeding a TF-IDF with topics of interest is a novel technique.

Table 1: Seed Topic descriptions and the number of sentences defining each topic.

Topic	Definition	Sentences
Confidence	**self efficacy, self perception of knowledge, ability**: Statements that express how an intern feels about themselves and their abilities, whether negative (unsure) or positive (sure). An expression of knowing (or not knowing) what to do and feeling that they are (or are not) grasping the material and new routines.	23
Overwhelmed	**stress, well being, burnout, anxiety**: Statements that express if an intern is bombarded with too many tasks or responsibilities and are having a hard time managing their time or specific situations, irrespective of the system. In some ways it could be the opposite of feeling confident, but one can be confident in your abilities and still be overwhelmed.	12
Supportive Environment	Statements about others around the intern and whether or not they can count on them for support.	14
System Issues	Statements that describe (positive or negative) issues related to technical systems (e.g., CERNER or EMR).	8

2.3 Annotation of AI Corpora

Two domain experts performed annotation on the Development and Test Corpora to classify each sentence as conveying one or none of the topics listed in Table 1. Both experts have a background in medical and clinical education, and one is a practicing physician. Annotation of the Development Corpus was done in 2 rounds, and included the document identifiers with ordered sentences. The first round had annotators independently annotate each sentence, then annotators came together to discuss a subset of annotations they disagreed on (15 sentences) and come to a consensus on the definition of each category. Then they independently re-annotated all sentences they disagreed on from round 1 (81 sentences). Disagreements in round 2 annotations were resolved by a third, non-medical annotator. If all three annotators classified a sentence differently, the sentence was set to "NA" (8 in Development Corpus, 12 in Test Corpus). The Test Corpus was annotated in one round with sentences randomly ordered and document identifiers removed (see Discussion) using the three annotations, with priority given to the two domain experts. Inter-annotator agreement was calculated as Cohens Kappa statistic. For round 1 of the development corpus the inter-annotator agreement between the two expert annotators was 0.35, which was then improved to 0.64 in round 2 (Table 2, top). The Test Corpus obtained a Kappa of 0.52 between the two expert annotators (Table 2, bottom).

Table 2: Inter-Annotator Agreement (Cohen's Kappa)

Development Corpus	All Seed Topics	Confidence & Supportive Environment Only
Expert 1 vs Expert 2	0.64	0.73
Expert 1 vs Annotator 3	0.48	0.71
Expert 2 vs Annotator 3	0.57	0.70

Test Corpus	All Seed Topics	Confidence & Supportive Environment Only
Expert 1 vs Expert 2	0.52	0.65
Expert 1 vs Annotator 3	0.46	0.75
Expert 2 vs Annotator 3	0.56	0.69

2.4 NLP Pipeline

The NLP pipeline to identify challenges expressed in reflective writing from AI students is composed of three primary steps as conveyed in Figure 1: Preprocessing, Sentence Representation, and Sentence Clustering. The pipeline assumes

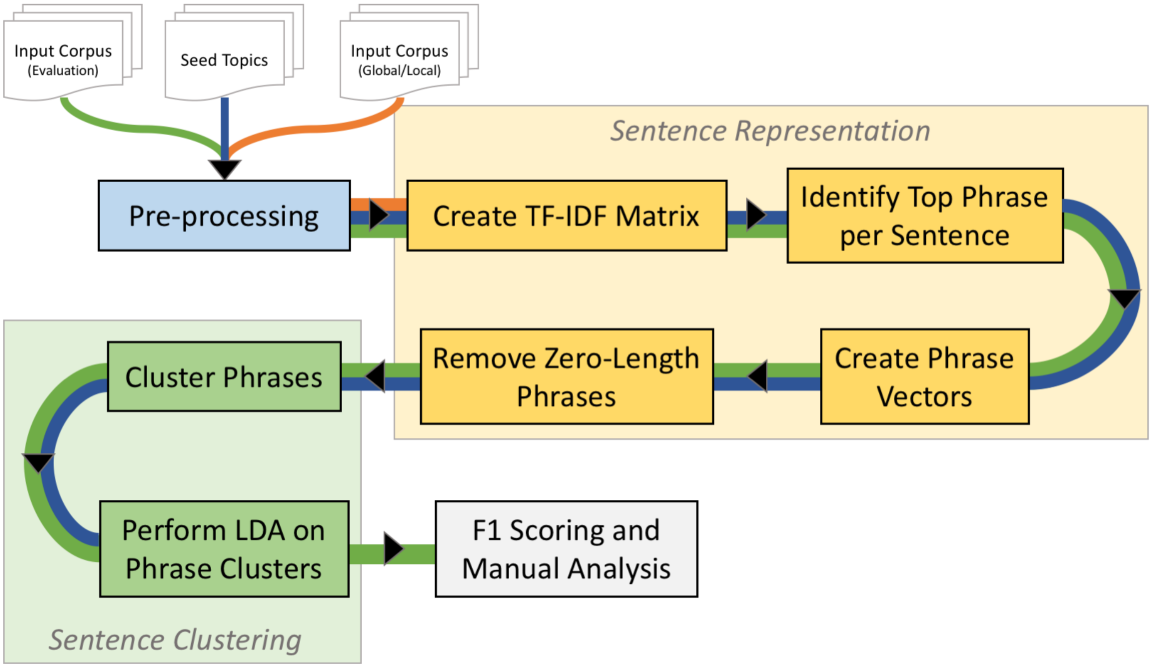

Figure 1: NLP Pipeline to identify Acting Intern challenges from reflective writing. Colored tracks indicate in which steps the various input corpora are utilized.

each sentence expresses one topic, which is extracted as a phrase, converted to a numerical array, and used to represent the sentence for clustering. These steps are now discussed in more detail.

2.4.1 TF-IDF Input Corpus

The full AI Corpus of 665 reflective writings is utilized to 1) obtain annotated subsets for evaluation and 2) as an input corpus to generate the TF-IDF matrix. The Development and Test Corpora discussed previously are the annotated subsets that will be processed and evaluated. In this section, two types of input corpora are utilized to generate the TF-IDF matrix: Global and Local. The Global input corpus refers to all 665 reflective writings in the AI Corpus plus the seed topics, whereas a Local input corpus only includes documents in the respective evaluation corpus plus seed topics as input to generate the TF-IDF matrix. For example, if we are processing the Test Corpus and want to use a Local TF-IDF matrix, then only the documents in the Test Corpus and the seed topics would be utilized to generate the TF-IDF matrix. If the Development Corpus is being processed, then the Local input corpus would include only documents in that corpus plus the seed topics. On the other hand, if we wanted the TF-IDF matrix to include more information we would utilize a Global input corpus that includes all 665 writings plus the seed topics to generate the TF-IDF matrix, which is used in the sentence representation step when processing the evaluation corpora. The following sections use the terms “Local TF-IDF” and “Global TF-IDF” to identify which type of input corpus was used to create the TF-IDF matrix.

2.4.2 Preprocessing

Each document’s text in the input and seed topic corpora is preprocessed using the Natural Language Toolkit (NLTK) package in Python[17]. First, sentences are split using the sentence tokenizer. Then common contractions are expanded prior to each sentence being tokenized by white space. Next, tokens are tagged with their part of speech before being lemmatized and lower-cased. Finally, stop words are removed, as well as all terms consisting of only numbers or punctuation. Stop words consist of those common English stop words defined by NLTK and a custom list of commonly used medical terms used by interns, such as ”patient”, that don’t provide much information in this context. The list of remaining terms for each sentence are utilized in downstream analyses.

2.4.3 Sentence Representation

In reflective writing, sentences can be long, rambling, and contain multiple topics; thus, in this work each sentence is limited to one topic by identifying its Most Informative Phrase (MIP). To identify each sentence’s MIP, documents in the input corpora, including seed topics and either the entire AI Corpus (Global TF-IDF) or a local subset (Local TF-IDF, i.e. the Development or Test Corpus), are first converted to a bag-of-words and serialized for input into Gensim[18] to calculate the NxM TF-IDF matrix, where N is the number of terms and M is the number of documents in the input corpora. Next, a document summarization method[†] is used at the sentence level to identify the MIP for each sentence using the calculated TF-IDF matrix. Each sentence is scanned using a window of 6 consecutive tokens. For each window, the MIP score is calculated using Equation 1 where i and j are the starting and ending term indices of the phrase window, respectively, $tfidf_m$ is the TF-IDF score for term m in the given document, p_m is a binary indicator for term m specifying if a term is an adjective/adverb ($p_m = 1$) or not ($p_m = 0$), and $sentiment_{i,j}$ is the sentiment polarity as calculated by TextBlob[§] for the current phrase.

$$MIPscore = \sum_{m=i}^{j}(tfidf_m * 3^{p_m}) * (1 + abs(sentiment_{i,j})) \quad (1)$$

The MIP score therefore puts an emphasis on terms that are describing words (adjectives and adverbs), and gives more weight to phrases with some type of positive or negative sentiment. Phrases with these properties should provide more information about how an intern is feeling rather than just conveying information. The phrase with the highest MIP score is chosen to represent the sentence. If the sentence is too short such that no 6-term phrase exists, then the sentence is represented by a zero-length phrase and removed prior to clustering. This rule results in the removal of some annotated sentences from clustering, which affects Recall calculations. Along with the raw recall results, the percentage of the maximum possible recall is also reported to assess how close a recall measure is to the maximum possible for a corpus using a 6-term phrase. Finally, the TF-IDF matrix is compressed using Singular Value Decomposition (SVD) and MIPs are converted to numerical vectors by summing the SVD transformed TF-IDF term vectors for each term in a MIP, which represents the distribution of a sentence across all corpus documents.

2.4.4 Sentence Clustering and Topic Analysis

The MIP vectors are used to cluster sentences across the input corpus by calculating one minus the cosine similarity between all pairs of vectors to create a dissimilarity matrix that is input into SciPy’s[19] scipy.cluster hierarchical agglomerative clustering algorithm and clustered using Ward linkage. The clustered dendrogram is cut at the highest threshold that results in all seed topics being assigned to different clusters. Finally, LDA analysis is perform on each cluster using Gensim to identify the main topic of each sentence cluster. The resulting clusters and associated topics are output into an Excel file for manual assessment.

2.5 Evaluation

Evaluation is performed by calculating Precision, Recall, and F1 measures for each seed topic independently using only sentences from the input corpus that were annotated with one of the four topics (sentences from seed topic documents are excluded). Precision is the ratio between correctly predicted mentions over the total set of predicted mentions for a specific entity; recall is the ratio of predicted mentions over the actual number of mentions; and F_1 is the harmonic mean between precision and recall.

[†] Method inspired by Charlie Greenbacker’s GitHub page at https://github.com/charlieg/A-Smattering-of-NLP-in-Python

[§] https://github.com/sloria/textblob

The underlying idea is that all sentences associated with the same topic should cluster together. For example, all sentences discussing System Issues should be located in the same cluster. In order to calculate the performance metrics, each seed topic is assigned to a cluster automatically by identifying the cluster with the highest number of true positive sentences from the given topic. Ideally, each seed topic should be assigned to a different cluster; thus, the dendrogram threshold is chosen to ensure this is the case.

Table 3: NLP pipeline results on Development and Test Corpora. For Recall, R is the obtained recall while R* refers to percentage of the maximum possible recall (see Discussion).

Corpus	TF-IDF	Clusters	Seed Topic	P	R (R*)	F1
Development	Global	25	Confident	0.80	0.17 (18%)	0.29
			Overwhelmed	0.75	0.25 (30%)	0.38
			Support Env	0.80	0.29 (34%)	0.42
			System Issues	1.0	0.63 (68%)	0.77
	Local	6	Confident	0.92	0.96 (100%)	0.94
			Overwhelmed	1.0	0.75 (90%)	0.86
			Support Env	1.0	0.79 (92%)	0.88
			System Issues	1.0	0.88 (96%)	0.93
Test	Global	18	Confident	0.83	0.28 (33%)	0.42
			Overwhelmed	0.50	0.25 (29%)	0.33
			Support Env	1.0	0.46 (58%)	0.63
			System Issues	0.67	1.0 (100%)	0.80
	Local	16	Confident	0.80	0.22 (26%)	0.35
			Overwhelmed	0.33	0.25 (29%)	0.286
			Support Env	1.0	0.15 (19%)	0.27
			System Issues	0.40	1.0 (100%)	0.57

3 Results

3.1 Development Corpus Performance

For the Development Corpus, the NLP pipeline achieved high precision for both the Global and Local TF-IDF matrices, with the lowest precision being 0.75 for the Overwhelmed seed topic. Recall differed between Global and Local with the Local TF-IDF matrix achieving much higher recall (over 90% of the maximum) than the Global across all seed topics (Table 3, top). This resulted in low Global F1 scores of less than 0.45 except for the System Issues seed topic, which had an F1 of 0.77. When using the Local TF-IDF matrix, F1 scores were above 0.86 for all seed topics. High performance is expected when using the Local TF-IDF matrix on the Development Corpus as the seed topics are generated from it, so the same documents are essentially being double counted and clustered. Reviewing the sentences in each cluster confirmed that seed topic sentences did cluster with the original document sentences to form cohesive topic clusters.

3.2 Test Corpus Performance

The Test Corpus was used to determine if seed topics generated from one set of challenge responses could identify sentences discussing similar challenges from another set of responses. When using either Global or Local TF-IDF matrices, the Confident and Supportive Environment seed topics obtained high precision (>0.8 and 1.0, respectively), while the Overwhelmed and System Issues received lower precision scores (Table 3, bottom). Interestingly, for the Confident/Supportive Environment seed topics the Precision stayed about the same from Global to Local, but the Recall dropped from 33%/58% to 26%/19%, respectively, whereas for the Overwhelmed/System Issues seed topics it reversed where the Recall stayed the same but the Precision dropped from 0.50/0.67 to 0.33/0.40, respectively. This indicates that the Confident/Supportive Environment clusters are well defined but we are not identifying all of the true positive instances when using the Local TF-IDF matrix. On the other hand, the Overwhelmed/System Issues clusters do identify the same number of true positive instances, but also pull in additional sentences as false positives, indicating that these topics may not be as well defined as the Confident/Supportive Environment seed topics. This could be due to the smaller number of sentences in the Overwhelmed/System Issues definitions, which may not be enough content to influence the TF-IDF matrix.

3.2.1 Seed Topic Analysis

While the performance of the NLP pipeline, as measured by F1 scores, is generally worse when using the Local TF-IDF matrix, the top terms for each of the clusters seem to be more relevant to the representative seed topic (Table 4). The Confident cluster, for example, includes words like "ive", "wasnt", and "right" along with relevant words like "confident" when using the Global matrix, whereas the Local TF-IDF matrix does not include these irrelevant terms and instead includes terms like "medication" and "shift". Inspection of the Confident cluster generated using the Local matrix reveals students had trouble with medication familiarity, adjusting to shift changes, and having confidence in their knowledge to diagnose a patient and come up with a treatment plan–all of which are relevant to ones confidence in themselves. In contrast, the Global cluster contained many more sentences that were irrelevant to having confidence, such as wishing colleagues good luck, being fortunate for having good people to work with, and getting sick during rounds. All these sentences seemed to be related by the key term "em", which is capitalized in the texts and stands for "Emergency Medicine". Utilizing the Global TF-IDF matrix must have put more weight on "em" because it appears frequently in a subset of student writings across the entire AI Corpus as the location of their AI, whereas it may not stand out as much in a local block of challenge responses.

Another example of how the Local TF-IDF matrix extracts more coherent seed topic clusters is with System Issues (Table 4). The System Issues topic did not appear in the Test Corpus as frequently as in the Development Corpus with only 2 sentences from the same blog post referring to EMR access issues. Thus, it makes sense that this cluster may contain more false positives or be very small. Interrogation of the cluster generated by the Global TF-IDF matrix reveals that along with the two true positive system issue sentences it includes discussions on being prepared for the next day, resident confusion, and not being able to get updates on patients. However, the cluster generated using the Local TF-IDF matrix is much more focused on students having trouble getting updates on their patients due to EMR access issues, communication issues, and not being able to put in or check on patient orders. Thus, while the Global obtained better precision, it was only because it included fewer other sentences, which were not very related to each other. On the other hand, the Local cluster contained more technical false positives, but these false positives were more cohesive and created a new topic surrounding obtaining timely updates on a patient's status.

Table 4: Keywords for LDA cluster topics for the Test Corpus.

TF-IDF	Seed Topic	Topic Keywords
Global	Confident	ive, em, attending, plan, confident, wasnt, think, diagnosis, right, found
	Overwhelmed	time, understand, vent, definitely, felt, head, responsibility, sedation, lot, learned
	Support Env	great, rotation, team, trying, think, picu, far, member, ive, helped
	System Issues	difficult, dont, resident, early, work, night, access, emr, sometimes, lot
Local	Confident	diagnosis, attending, shift, thinking, plan, ruling, medication, confident, think, hardest
	Overwhelmed	time, better, stressor, presentation, feel, seen, definitely, felt, head, taking
	Support Env	different, experience, great, another, em, far, challenge, currently, rotation, team
	System Issues	update, dont, order, need, frequently, made, difficult, forward, going, work

4 Discussion

4.1 Annotation Challenges

Annotating the reflective writings from medical students was difficult due to the subjective nature of the content. Prior to the round 1 annotation of the development corpus, annotators had a brief discussion of the seed topics; however, annotations suggested a consensus had not been reached. This was evident in the annotations for System Issues where one annotator defined this strictly as technical system problems; however, the second annotator had a broader definition that included educational and organizational system issues. After meeting to discuss definitions, there was still a discrepancy in how System Issues was annotated, which resulted in lower inter-annotator kappa values when using all topics. For this work, the definition of System Issues is limited to technical system issues as defined in Table 1. All non-technical system issue annotations were set to "NA" in the gold standard, but were left as-is for the inter-annotator calculation.

Another difficulty in annotating these reflective writings was keeping the annotations to the sentence level. For the first round of annotations, annotators were given sentences in the same order as they appeared in the student responses. This allowed them to utilize the document context to annotate sentences even though a sentence by itself did not express the seed topic it was annotated with. The Overwhelmed seed topic was the major culprit in this instance where several sentences read together can express a feeling of being overwhelmed, but individually the sentences don't necessarily convey this feeling. Thus, for the annotation of the Test Corpus, all sentences were randomly ordered and document identifiers were removed. However, a large discrepancy in which sentences convey being overwhelmed still exist in the final annotations, which lead to low Kappa values.

The difficulty in annotating this corpus demonstrates how subjective interpreting reflective writing can be, which significantly contributes to the difficulties in developing an automated NLP pipeline. However, while System Issues and Overwhelmed topics were ambiguous, there was a fair amount of agreement on sentences discussing Confidence and Supportive Environment. When the System Issues and Overwhelmed seed topics are removed from the inter-annotator agreement calculation the Kappa scores increases to approximately 0.70 (Table 2, bottom) for both the Development and Test Corpus. This result indicates that there are some topics with a common definition that may be able to be identified utilizing an automated approach along with new challenges experienced by the medical students.

4.2 Local vs Global TF-IDF

While F1 performance was poor when using a Local corpus to generate the TF-IDF matrix, the sentences in a seed topic cluster seemed more related to each other as compared to the seed topic clusters generated using the Global TF-IDF matrix. Normally, in NLP one would think more data is better; however, with this corpus the empirical results indicate that more information just adds noise. This result might be attributed to the organization and collection format of the corpus where reflective writings are extracted from forum post threads where all responses within a thread are not independent. That is, when an intern posts their answer to the question they are able to see all prior responses, and their writing and selection of topics to discuss may be influenced by what others have already written. This can be seen in the frequent sentences indicating they agree with previous posts and that they enjoyed reading everyone's thoughts. Thus, for this type of prompted, block structured corpus, we found that utilizing only the data within each block was able to better identify known challenges experienced by the students. While the use of a Local TF-IDF matrix resulted in more cohesive seed topic clusters, it failed to identify several new topics that appeared when using the Global TF-IDF. Therefore, depending on the primary objective, the best type of TF-IDF matrix utilized may differ.

4.3 Identifying New Challenges

In this work, we had the advantage of knowing which sentences belonged to which topics; however, when executing this method on an unannotated corpus it is helpful to know how to identify clusters that potentially contain new challenges. For the Development and Test Corpora used in this study we noticed that cohesive topics came from clusters that included sentences from multiple student responses. For example, the Confident and Overwhelmed clusters contained sentences from 5 students each when using the Local TF-IDF, Supportive Environment contains sentences from 7 students, and System Issues contains sentences from 3 students. The remainder of the clusters in the Local results contained responses from only 1 or 2 students. However, in the Global results there were several other clusters with sentences from multiple students that identified common challenges in addition to the seed topics. These included clusters describing difficulties in finding ones place on a team, difficulties staying updated on assigned patients, and the challenge of adapting to differences in specialties, devices, and surgeon preferences. In the future, these block-level themes can be extracted from multiple blocks to identify more global challenges experienced by all AI students.

5 Limitations

The subjectivity of interpreting reflective writing texts is a major hurdle for anyone processing this type of data, and is demonstrated in this work by the disagreement of the annotators as to how to assign sentences to seed topics. This low annotator agreement makes quantitative evaluation of results difficult. To help mitigate this issue in the future, annotators will discuss and agree on more specific definitions of seed topics, and will be given practice data sets that have been vetted thoroughly by multiple experts for training purposes. Another limitation of this work includes not being able to utilize the context of neighboring sentences in the NLP pipeline. In the AI reflective writings, concepts, such as feeling overwhelmed, can be relayed over multiple sentences. Having the ability to take neighboring context into account may aid in identifying more cohesive challenges. In addition, this method currently only allows one phrase (or topic) per sentence even though there may be multiple topics discussed in the same sentence. This was observed in the Development Corpus where the first part of a sentence discussed having confidence and the last part of the sentence talked about also being overwhelmed. Having the ability to break up these compound sentences will aid in challenge identification.

6 Conclusions and Future Work

In conclusion, we have presented a novel approach to identify common themes from reflective medical writings by utilizing a seeded TF-IDF matrix, which emphasizes known topics while also allowing new topics to emerge. Clustering sentence vectors created from the seeded TF-IDF produces useful topic clusters that are easily and quickly digested by medical educators compared to reading and manually categorizing the hundreds of responses they collect each year. This information is utilized to identify and prioritize areas in the Acting Internship program that should be altered or updated to provide medical students with the tools and knowledge they need to successfully complete their AI and transition from a medical student to a practicing intern.

Future work will include examining the potential of expanding the seed topic dictionary, as with each new cohesive cluster a new seed topic can be generated. Additionally, through the annotation process the domain experts identified multiple types of system challenges. While we focused on technical system challenges in this work, students also expressed difficulties in navigating educational and organizational systems that will be useful to identify. Since these other system challenge concepts have a different lexicon they will need their own seed topic defined and cannot be combined with the current technical system challenge concept. The Test Corpus in this work focused on a single AI block, however, dividing the challenge responses up differently may provide additional insights into challenges experienced by students. For example, processing responses by department may provide insight on department-specific challenges not experienced elsewhere, or analyzing each year to identify how challenges change over time.

Finally, the implementation of this system is domain agnostic in that any corpus can be used to create and/or seed the TF-IDF matrix. Thus, it can be used for other tasks that require the extraction of common themes from a text corpus. The implementation of using a local TF-IDF matrix would benefit the analysis of corpora that have a niche lexicon, such as medical writings or reports. For example, identifying common themes discussed in clinician-patient narratives is another application that could benefit from this method. Currently, this pipeline includes manual steps, such as data cleaning, and requires some knowledge of coding to execute, which limits its portability to other departments or institutions. We are currently working to fully automate this pipeline and develop a graphical user interface with visualization of the results to make this tool available for both medical educators outside our institution and others to extract common themes from corpora with a niche lexicon.

References

[1] Tracy Moniz, Shannon Arntfield, Kristina Miller, Lorelei Lingard, Chris Watling, and Glenn Regehr. Considerations in the use of reflective writing for student assessment: issues of reliability and validity. *Medical Education*, 49(9):901–908, September 2015.

[2] John Sandars. The use of reflection in medical education: AMEE Guide No. 44. *Medical Teacher*, 31(8):685–695, January 2009.

[3] Emma-Jane Berridge, Della Freeth, Judi Sharpe, and C. Michael Roberts. Bridging the gap: supporting the tran-

sition from medical student to practising doctor a two-week preparation programme after graduation. *Medical Teacher*, 29(2-3):119–127, January 2007.

[4] Doug Franzen, Amanda Kost, and Christopher Knight. Mind the gap: the bumpy transition from medical school to residency. *Journal of Graduate Medical Education*, 7(4):678–680, December 2015.

[5] Faizal A. Haji, David B. Clarke, Marie C. Matte, David M. Brandman, Susan Brien, Sandrine de Ribaupierre, Cian OKelly, Sean Christie, Patrick J. McDonald, Abhaya V. Kulkarni, Simon Walling, and Anna MacLeod. Teaching for the transition: the Canadian PGY-1 neurosurgery Rookie Camp. *Canadian Journal of Neurological Sciences*, 42(1):25–33, January 2015.

[6] Alan R. Teo, Elizabeth Harleman, Patricia S. OSullivan, and John Maa. The key role of a transition course in preparing medical students for internship. *Academic medicine : journal of the Association of American Medical Colleges*, 86(7):860–865, July 2011.

[7] Janet E. Dyment and Timothy S. O'Connell. Assessing the quality of reflection in student journals: a review of the research. *Teaching in Higher Education*, 16(1):81–97, February 2011.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[9] Ye Chen, Bei Yu, Xuewei Zhang, and Yihan Yu. Topic modeling for evaluating students' reflective writing: a case study of pre-service teachers' journals. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, LAK '16, pages 1–5, New York, NY, USA, 2016. ACM.

[10] Andrew Gibson and Kirsty Kitto. Analysing reflective text for learning analytics: an approach using anomaly recontextualisation. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, LAK '15, pages 275–279, New York, NY, USA, 2015. ACM. event-place: Poughkeepsie, New York.

[11] S. R. Vrana, R. S. Bono, A. Konig, and G. C. Scalzo. Assessing the coherence of narratives of traumatic events with latent semantic analysis. *Psychological trauma : theory, research, practice and policy*, October 2018.

[12] Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284, January 1998.

[13] Leonard K. M. Poon, Zichao Li, and Gary Cheng. Topic classification on short reflective writings for monitoring students progress. In Simon K.S. Cheung, Lam-for Kwok, Will W.K. Ma, Lap-Kei Lee, and Harrison Yang, editors, *Blended Learning. New Challenges and Innovative Practices*, Lecture Notes in Computer Science, pages 236–246. Springer International Publishing, 2017.

[14] Richard Schwartz, Sreenivasa Sista, and Timothy Leek. Unsupervised topic discovery. In *Proceedings of Workshop on Language Modeling and Information Retrieval*, 2001.

[15] J. Jayabharathy, S. Kanmani, and A. A. Parveen. Document clustering and topic discovery based on semantic similarity in scientific literature. In *2011 IEEE 3rd International Conference on Communication Software and Networks*, pages 425–429, May 2011.

[16] Thomas Roelleke and Jun Wang. TF-IDF uncovered: a study of theories and probabilities. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 435–442, New York, NY, USA, 2008. ACM. event-place: Singapore, Singapore.

[17] Steven Bird, Edward Loper, and Ewan Klein. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

[18] Radim Rehurek and Petr Sojka. Software framework for topic modelling with large corpora. In *In Proceedings of the Lrec 2010 Workshop on New Challenges for Nlp Frameworks*, pages 45–50, 2010.

[19] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001.