Using Text Analytics on Reflective Journaling to Identify Cultural Capitals for STEM Students

Karuna Nayak*[1], Shailesh Krishna†, Khanh Tran‡§, Marissa Harris§, Ana Maria Barrera^x, Kim Coble‖, Anagha Kulkarni¶[2]

*Hewlett Packard Enterprise, †Dept. of Decision Sciences and Information Systems, ‡Dept. of Ethnic Studies, §Dept. of Biology, ^xDept. of Kinesiology, ‖Dept. of Physics and Astronomy ¶Dept. of Computer Science, San Francisco State University, California 94132, U.S.A.

karuna.nayak@hpe.com, {skrishna, ktran1, mlharris}@mail.sfsu.edu, {abarrera, kcoble, ak}@sfsu.edu

***Abstract*—Alma project focuses on retention of Historically Underrepresented (HU) students in STEM fields by affirming their cultural capitals (skills, knowledge, communities, and abilities that students possess). Students are asked to write reflective journals on prompts such as 'Why am I here?' to reminds them of their cultural capitals, and career aspirations. These essays are then analyzed by the Alma team to identify mentions of cultural capital themes (CCT) such as Attainment (describes a tangible goal), or First Generation (identifies student being first in family to attend college). However, detecting these CCTs in student essays is a labor intensive task. Our work offers a computational solution that is scalable and effective for this problem. We propose TACCTI (pronounced as *Tacit*), Text analysis and mAchine learning for Cultural Capital Theme Identification – a computational framework that employs latest developments in natural language processing and machine learning to identify instances of cultural capital themes in student essays. An empirical evaluation using Logistic Regression, Random Forest, LSTM, and BERT algorithms for the task of detecting presence (or absence) of Attainment CCT provides F1 scores of 0.82, 0.81, 0.84, and 0.92, respectively. Classification models for detecting First Generation CCT demonstrate that when limited training data is available, the traditional classification algorithms outperform deep learning based approaches – Logistic Regression and BERT models offer effectiveness of 0.91 and 0.83, respectively. Additional analysis demonstrates that word embeddings and named entity type of features (signals) have the highest influence on the performance of the classification models.**

***Index Terms*—Educational Text Analytics, Community Cultural Wealth, Reflective Journal Analysis, BERT**

I. Introduction

Historically Underrepresented (HU) students in STEM fields (e.g., African American, Hispanic or Latino/Latina, American Indian, and Alaskan Natives) often come from cultural backgrounds that are different from majority of the students in the field [4], [6]. Their unique backgrounds and cultures can be source of strength that helps HU students persist through the challenging STEM programs. This is the motivation behind a novel program, Alma, implemented at San Francisco State University (SFSU). Alma means heart/soul in Spanish, and the Alma program aims to support HU students in STEM programs by affirming their cultural wealth and sense of belonging. Cultural wealth can be defined as the skills, knowledge and abilities possessed and used by certain communities [4]. Social and clinical psychology research indicates that writing about any major life events helps in physical and mental health with the help of self-affirmation and cognitive processing [5]. As part of the Alma program students are asked to reflect and respond to essay prompts such as 'why am I here?', 'what do I do when life gets challenging' – prompts that are designed to help students remember and assert their strengths, career motivations, life purpose and experience. Students are also encouraged to discuss their responses and life experiences with their peers to promote community development and social inclusion – activities that have shown positive impact on HU students' retention in STEM fields [1]–[3].

Early on in the Alma program, the students' essay responses were analyzed by a team of 11 researchers to identify the mentions of cultural values and assets (i.e. cultural capital) that students leverage. Alma team identified 11 different cultural capital themes (CCTs) in student essays. Some of the CCTs are *Attainment*: mention of tangible goal(s) (i.e. something that could be added to a CV or resume), *Community Consciousness*: mention of solidarity with community and the desire to give back to a community one identifies as being part of, *Familial*: mention of support provided by family, whether tangible support (e.g. food, financial support), emotional support, or role modeling, and *First Generation*: mention of being the first in their family to attend college. Each essay was analyzed by 2-3 researchers. For those essays on which the cultural capital themes differed between researchers, the selections were negotiated until 100% agreement was reached on the final codes to be assigned. These themes are inspired by the rich body of work on community cultural wealth frameworks that establish the importance of using asset-based interventions rather than deficit-based interventions for promoting students' academic progress and success [4], [6], [14]–[16]. The goal of the thematic analysis of student essays is to help educators

[1]Work done while at SFSU.

[2]To whom correspondence should be addressed.

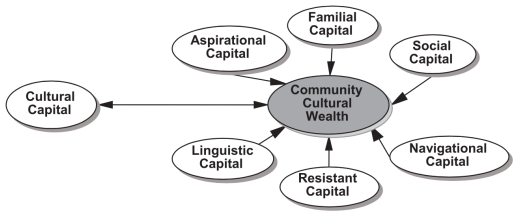

Fig. 1. Cultural Wealth Model by T.J.Yasso [4]

understand and leverage the strengths of their students to foster their success in the field.

Unfortunately, identifying cultural capital themes in student essays is a labor-intensive task that is costly in time and effort. As a result, it does not scale to large class sizes and institution-level broad implementation. To address this challenge, we propose TACCTI (pronounced as *Tacit*), Text analysis and mAchine learning for Cultural Capital Theme Identification – a scalable computational framework that employs latest developments in natural language processing and machine learning to identify instances of cultural capital themes in student essays. We focus on two specific CCTs in this work, *Attainment* (e.g. *I want to get a degree in Biology*), and *First-Generation* (e.g. *I'm the first in my family to go to college*).

II. Related Work & Background

The cultural capital themes in the Alma project are inspired from the community wealth model [4], [6]. Author Tarra J. Yasso explains the Critical Race Theory (CRT) in [4] as a framework which could be used to examine and challenge the direct or indirect impact of race and racism on our social structure and practices. The theoretical framework of the CRT traditionally interprets that the students of color brings the cultural deficiencies. Author Yasso introduce an alternative concept, Community Cultural Wealth and discusses the potential of cultural wealth brought by students of color into schooling process. The author defines 6 forms of cultural capitals such as aspirational, navigational, social, linguistic, familial, and resistant capital as shown in Figure 1. Alma team performed an iterative process of identifying cultural capitals which better reflect the experience of SFSU students in STEM courses and extended Yasso's cultural wealth model further for additional cultural capitals.

Text classification is one of the classic problems in Natural Language Processing. Due to the recent advancement of technologies, deep learning models are widely used for classification tasks. In [17], Marwan et al. implements the convolutional neural network (CNN), bidirectional long short-term memory (LSTM) and bidirectional gated recurrent units (GRU) models for toxic comments classification with imbalanced data. They compare the model results with an ensemble model which utilizes CNN, LSTM and GRU models and ensemble model performs better. Devlin et al. proposed BERT (Bidirectional Encoder Representations from Transformers) to train deep bidirectional representations from unlabeled texts [9]. The BERT architecture obtains the state-of-the-art results on many NLP tasks. Sun et al. discusses the BERT fine-tuning techniques for text classification, and BERT performance for longer text, layer selection and learning rates [18].

Model	Naïve Bayes		Logistic Regression	
	Essay	Sentence	Essay	Sentence
F1-Score	0.42	**0.65**	0.57	**0.7**

TABLE I
Classification model performance for essay-level dataset Vs sentence-level dataset

III. TACCTI (pronounced as *Tacit*)

TACCTI is a framework for identifying instances of cultural capital themes (CCTs) in student essays that are written as part of self-affirming and reflective journaling exercises. Currently TACCTI supports identification of two CCTs (*Attainment* and *First-generation*) but the framework is designed to easily onboard additional CCTs as their identification models become available. A high-level system architecture of TACCTI is depicted in Figure 2.

A. Input Data Pre-processing (IDP)

The first component of the system, Input Data Pre-processing (IDP), was born out of one of the early design decisions we had to make: the granularity of operation – would CCT identification be conducted at essay-level or essay sentence-level?

1) Granularity of operation: The educator is typically interested in knowing which essays contain instances of CCTs, not which essay sentences contain instances of CCTs. However, only a few sentences in a given essay contain CCTs, thus operating at essay-level might dilute the CCT signal. To test this hypothesis we conducted a preliminary experiment for *Attainment* CCT identification at both granularity levels.

Two commonly used classification algorithms, Naïve Bayes and Logistic Regression, were used to train separate classification models with essay-level data and sentence-level data. Traditionally used feature types, such as Part Of Speech (POS) tags, unigrams, sentiment polarity and subjectivity scores were employed for both essay- and sentence-level classification models. The dataset used for this experiment consisted of 593 essays for the prompt *"why am I here"*. On average, the essays consisted of 15–20 sentences (796 words), and when present, the *Attainment* theme is mentioned in only 1–2 sentences of the essay. Thus there are many more sentences that do not contain *Attainment* theme (Class 0) than sentences with *Attainment* theme (Class 1). This imbalance in class distribution is also present in essay-level data but it is less skewed as is visualized in Figure 3. Sentence-level dataset has 14% (Class 1) : 86% (Class 0) class distribution, while essay-level dataset is 28% (Class 1) : 72% (Class 0) class distribution. These datasets are divided into 80:20 train:test splits for model training and evaluation, respectively. Data split is done with stratification to maintain the same class distribution in both, train and test splits. Classification results

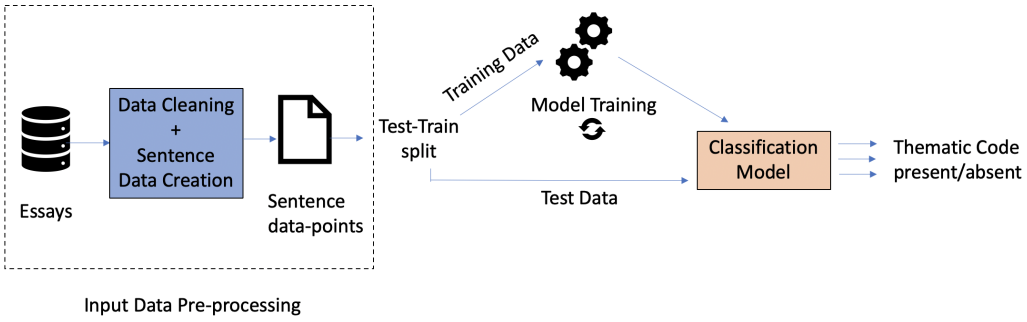

Fig. 2. TACCTI System Architecture

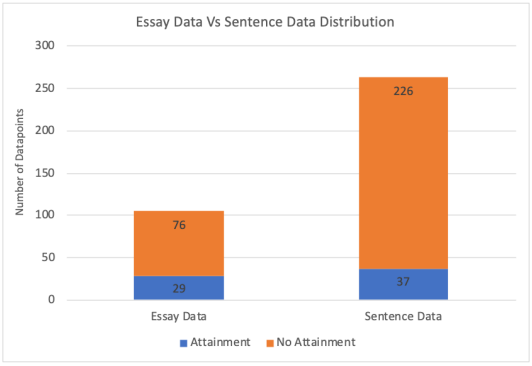

Fig. 3. Class distribution for essay-level data and sentence-level data

for the four models (2 algorithms x 2 dataset granularity levels) on the test split are summarized using F1-score in Table I. For both algorithms, Naïve Bayes and Logistic Regression, the sentence-level dataset provides substantially better performance than essay-level dataset. This experiment informed our design decision to operate at sentence-level for the task of CCT identification in student essays.

2) Sentence-level Annotation: The decision to operate at sentence-level necessitated additional processing on the human annotated data because the Alma team had conducted the CCT annotation task at essay-level. Annotators were instructed to specify presence or absence of a specific CCT in a given essay, and to also provide an excerpt from the essay where the CCT was mentioned. These excerpts were often phrases; not complete sentences from the essay. Furthermore, many times the excerpts were not copied verbatim from the essays. Thus exact match could not be employed to identify the essay sentence that contained the excerpt. Also, in few essays, there were multiple sentences that contained the excerpt. To handle all of these scenarios, the following multi-step approach was designed and implemented to generate sentence-level annotated data from the original essay-level annotated data.
Input: Essay-level human annotated data (CCT present or absent label, and excerpt).
Output: Sentence-level (derived) annotated data.

1) Split the essay into sentences using full-stop '.' as sentence termination mark.
2) Check for the exact match of the excerpt with each sentence using regular expression. In case of exact match, annotate the sentence with Class label 1 (CCT present).
3) If there is no exact match with the excerpt, perform the following *NE-based match* on each sentence. Extract all the Named Entities (NEs) for job title and profession from the excerpt[1]. If all the extracted NEs are present in a sentence, annotate the sentence with Class label 1.
4) If NE-based match does not return any sentences, perform the following *NA-based match* on the sentence. Extract all types of Nouns (noun singular, noun plural, proper noun singular, proper noun plural) and Adjectives (adjective, adjective comparative, adjective superlative) from the excerpt using POS tagger. If all the extracted POS tags are present in the sentence as well, annotate the sentence with Class label 1 else with Class label 0 (CCT absent).

B. Classification Models: Attainment CCT Identification

We frame the task of identifying Attainment theme in student essay-sentences as a binary classification problem: given a student essay-sentence, is the Attainment theme present or absent. For this problem, we test multiple classification models that are progressively more complex and are described next.

1) Base Model: The base model for Attainment CCT identification in TACCTI is built using Logistic Regression algorithm for its computational efficiency and model interpretability. The set of features employed for this model consisted of unigrams with minimum document frequency of two, POS tag count, sentiment polarity, and subjectivity scores obtained using TextBlob [20]. The parameters for Logistic Regression model are set to liblinear optimization, L2 penalty, and max 100 iterations for convergence.

When the trained classification model was analyzed for feature importance, it was noted that the career and education related unigram features received higher weight. This aligns

[1]Stanford Core NLP NER module with the help of Pycorenlp wrapper is used for extracting NEs [10].

with the factors that the human annotators considered when annotating essays for Attainment CCT. However, the downside of having such unigrams as high weight features is the generalizability of the model. For example, if a model is trained using Biology student essays, leading to higher weights for unigrams such as *sonography*, *optometrist* and *cardiothoracic*, then this model is unlikely to perform well on student essays from other majors. To address this drawback additional features were designed to improve generalizability.

2) LR Model: Three additional feature categories are leveraged by this next model: (1) STEM similarity features, (2) Named Entity feature (3) Word embeddings.

STEM similarity features: The intuition behind STEM features is that presence of any STEM vocabulary provides a generic signal about potential presence of Attainment theme. To operationalize this intuition, Physics, Chemistry, Biology, Technology, Engineering and Mathematics synsets[2] are obtained from Wordnet [19]. Each sentence datapoint, is tokenized and synsets of these tokens (except stopwords) are obtained. Using Wordnet's *path similarity module*, a similarity score between token synsets and STEM field synsets is computed. This generates six STEM similarity features – one for each of the above STEM fields.

Specializations in medical field such as *optometric* and *cardiothoracic* does not exist in Wordnet. As a result, when student essays use these terms to indicate career goals, this Attainment signal could not be leveraged. To address this problem, a list of medical specializations[3] was compiled, and new feature, *MedSp*, is defined that counts the number of medical specializations present in the sentence.

Named Entity feature: When an essay-sentence mentions a profession or job title, it presents a strong signal for Attainment theme. To capture this observation, the *NE* feature is defined that counts the number of Named Entities of type profession or title in the sentence.

Word embeddings features: Word embeddings have emerged as the most effective technique to represent the various semantic concepts that a word can convey or invoke [12], [13]. In TACCTI, word embeddings were added as features to leverage their rich semantic representation power and to generalize the model better. Based on preliminary experiments that compared Glove Embedding [12] and Word2Vec embedding [13], 100-dimensional Word2Vec embeddings that were trained on student essays were chosen.

3) RF Model: This next Attainment CCT identification approach in TACCTI is built using Random Forest (RF) [7], a widely-used ensemble classification algorithm that is known to perform well in presence of noisy features and skewed class distribution. RF models are also easy to interpret. The RF classification model learnt in TACCTI for Attainment CCT identification builds 200 decision trees with maximum depth of 2. The choice of features for this model is guided by the *Feature Ablation* study described in Section V-B. The following features showed the largest contributions in the ablation study and thus were included in the RF model: POS tag count, NE feature, and word embeddings features. This subset of all features was chosen to avoid model overfitting since the amount of labelled data is limited.

4) LSTM Model: The next two models leverage developments in deep learning for Attainment CCT identification in TACCTI. Long-Short Term Memory (LSTM) networks are a type of Recurrent Neural Networks (RNN) which are capable of handling long term dependencies and are designed to overcome the problem of short term memory in traditional RNNs [21]. Since we are dealing with sentence-level language data, LSTM network is expected to perform better than a model with hand crafted features. This is because LSTM can regulate the information flow and can learn the sequence information to make better predictions. The bi-directional LSTM [8] understands the context better as it combines the hidden states to preserve information from both past and future at any point in time.

The neural network for Attainment classification consists of an embedding layer with 100-dimension Word2Vec embedding, 2 bi-directional LSTM layers with a dropout layer of 0.5 in between to avoid over-fitting of data and finally a dense layer with tanh activation function and a dense output layer with sigmoid activation function. The binary cross-entropy loss function is used for binary classification.

5) BERT Model: Bidirectional Encoder Representations from Transformers (BERT) is the latest advancement in language representation techniques that allows pre-trained model to be easily fine-tuned with task specific data [9]. Of the two BERT models made available by Devlin et al., we employ the BERT_{BASE} model which is pre-trained on large unlabeled general domain corpus, and consists of 12 transformer blocks, 12 self-attention heads, and the hidden size of 768. For training the Attainment CCT classification model, the pre-trained BERT_{BASE} model parameters are fine-tuned using the labeled sentence-level data. We have used Ktrain[4] wrapper built on top of Keras to load and fine-tune the BERT_{BASE} model[5].

C. Classification Models: First Generation CCT Identification

Identifying instances of First Generation CCT in student essay-sentences is comparatively easier than other thematic codes because often students explicitly state that they are first generation college going student in their family. However, the challenge is that the First Generation CCT does not get mentioned as often as Attainment CCT in student essays. This is so because not every student is the first in their family to go to college but most students do come to college with specific career goals. Thus, the labelled data for First Generation CCT is limited. As a result, when training the classification model with Logistic Regression for First Generation CCT, we employ

[2]Wordnet defines synsets as *Nouns, verbs, adjectives and adverbs that are grouped into sets of cognitive synonyms, each expressing a distinct concept* [19].

[3]https://www.sgu.edu/blog/medical/ultimate-list-of-medical-specialties

[4]https://github.com/amaiya/ktrain

[5]Using the learning rate finder of Ktrain module, optimal learning rate 2e-05 was set to train the model efficiently for 5 epochs.

	Precision	Recall	F1	DataPoints
Class 0	0.94	0.98	0.96	48
Class 1	0.67	0.40	0.50	5
Macro Average score	0.81	0.69	0.73	53

TABLE II
RESULTS WITH BASE MODEL FOR ATTAINMENT CCT CLASSIFICATION

	Precision	Recall	F1	DataPoints
Class 0	0.97	0.87	0.92	116
Class 1	0.61	0.88	0.72	26
Macro Average score	0.79	0.88	0.82	142

TABLE III
RESULTS WITH LR MODEL FOR ATTAINMENT CCT CLASSIFICATION

the simple feature set of unigrams to avoid overfitting the model. We also experiment with BERT$_{BASE}$ model for First Generation CCT classification since this approach provides the most effective model for Attainment CCT (Section III-B5). The pre-trained BERT$_{BASE}$ model is fine tuned using auto-fit function and learning rate of 2e-05, same as for Attainment model.

IV. EXPERIMENTAL METHODOLOGY

A. Dataset

Attainment CCT: Out of 593 annotated essays for prompt 'Why I am here', Attainment CCT was mentioned in 196 essays. These 196 annotated essays were converted into sentence-level dataset using the approach described in Section III-A2, which generated 1412 number of labelled datapoints (Class 1: 254 sentences (18%), Class 0: 1158 sentences (82%)). This data was then divided into 90:10 train and test splits which were used for training and evaluation of all Attainment classification models. The train and test splits maintain the same class distribution as the overall labelled data, 18% and 82%. The train split consists of 1270 datapoints (Class 1: 228 sentences, Class 0: 1042 sentences), and test split contains 142 datapoints (Class 1: 26 sentences, Class 0: 116 sentences).

First Generation CCT: The Firs Generation CCT was found in 42 essays during annotation. These essays were converted to sentence-level dataset of 357 labeled datapoints. This data was divided into 80:20 train and test splits with same class distribution of 14% and 86%. The train split consists of 285 datapoints (Class 1: 40 sentences, Class 0: 245 sentences), and test split contains 72 datapoints (Class 1: 10 sentences, Class 0: 62 sentences).

B. Evaluation Metrics

The TACCTI framework implements three widely used evaluation metrics to quantify and empirically evaluate the performance of classification models: Precision, Recall and F1-measure. Precision metric for a Class x computes the fraction of datapoints that are correctly predicted as Class x. Recall metric for Class x quantifies the fraction of true Class x datapoints that were classified correctly. F1-measure for Class x computes the harmonic mean of Precision and Recall for Class x. The Macro average score for each of the metrics provides a consolidate score across all classes, where every class receives an equal weight. This is important for setups such as ours where the class distribution is highly skewed and the smaller class is just as important as the dominant class.

V. RESULTS AND ANALYSIS

The first part of this section presents the results for all Attainment CCT classification models and feature ablation study conducted with LR model for Attainment CCT. The later part of the section discusses results for the First Generation CCT classification models.

A. Attainment CCT: Base Model

The Base model for Attainment CCT classification was built very early on in this project when only a subset of the labeled data was available, specifically, 105 annotated student essays (263 sentence-level datapoints with 14% Class 1 labels). The dataset is split into 80:20 Train and Test splits.

The results for Attainment CCT classification with Base Model on the test data are provided in Table II. Recall that the Base model employs Logistic Regression algorithm with feature set consisting of unigrams, POS tag count, sentiment polarity and subjectivity score (Section III-B1). As evidenced by the results, this model struggles with Class 1 (Attainment), which is the smaller class. The low Recall score for Class 1 indicates that large fraction of true Class 1 datapoints (60%) are misclassified as Class 0 by this model. This suggests that the learnt classification model is biased toward the larger class (Class 0), and that the employed features have very limited representational power.

B. Attainment CCT: LR Model

The next Attainment CCT classification model, LR Model, has additional feature types: STEM similarity, Named Entity feature, and Word embeddings (Section III-B2) to improve the data representation ability of the model. Table III provides the classification results for the LR Model. The utility of the additional features is evident – Class 1 performance is substantially higher for this model compared to Base Model. Naturally, the overall Macro averaged performance of this model is also higher than the previous model. The improvement in Recall for Class 1 indicates reduction in misclassification of true Class 1 datapoints. However, Precision for Class 1 lags, suggesting that this model is over-predicting Class 1 (misclassifies true Class 0 datapoints as Class 1). This trend is exactly opposite to the one observed with the previous model – now the learnt classification model seems to be biased toward the smaller class (Class 1). A potential reason for this could be that the number of features was disproportionately large as compared to the available training data. To test this hypothesis a few different subsets of feature types were experimented with. The goal was to check if any of the feature subsets provide same or comparable performance as the full feature set to identify the redundant features. Thus reducing the number of features in

Combination #	**Features**	**Precision**	**Recall**	**F1**
#1	POS tag+Embedding	0.77	0.91	0.80
#2	Named Entity+Embedding	0.78	0.86	0.81
#3	POS tag+Sentiment+Unigrams	0.77	0.82	0.79
#4	**POS tag+Named Entity+Embedding**	**0.79**	**0.88**	**0.82**
#5	Sentiment+STEM Similarity+Embedding	0.76	0.87	0.78
#6	STEM Similarity+Named Entity+Embedding	0.77	0.84	0.79
#7	Sentiment+STEM Similarity+Named Entity+Embedding	0.77	0.84	0.79
#8	STEM Similarity+POS tag+Named Entity+Embedding	0.79	0.86	0.81
#9	Sentiment+STEM Similarity+POS tag+Named Entity+Embedding	0.78	0.86	0.81
	All feature types	0.79	0.88	0.82

TABLE IV
RESULTS FROM FEATURE SUBSET STUDY WITH LR MODEL FOR ATTAINMENT CCT CLASSIFICATION

	Precision	Recall	F1	DataPoints
Class 0	0.97	0.86	0.91	116
Class 1	0.59	0.88	0.71	26
Macro Average score	0.78	0.87	0.81	142

TABLE V
RESULTS WITH RF MODEL FOR ATTAINMENT CCT CLASSIFICATION

	Precision	Recall	F1	DataPoints
Class 0	0.94	0.95	0.94	116
Class 1	0.76	0.73	0.75	26
Macro Average score	0.85	0.84	0.84	142

TABLE VI
RESULTS WITH LSTM MODEL FOR ATTAINMENT CCT CLASSIFICATION

	Precision	Recall	F1	DataPoints
Class 0	0.98	0.96	0.97	116
Class 1	0.83	0.92	0.87	26
Macro Average score	0.90	0.94	0.92	142

TABLE VII
RESULTS WITH BERT MODEL FOR ATTAINMENT CCT CLASSIFICATION

the model. The LR Model employs six feature types: unigrams, sentiment (sentiment and polarity scores), POS tag count, NE count (title and profession), STEM similarity, and Word embeddings. The results from the feature subset study are in Table IV. The Macro average scores for Precision, Recall, and F1 are provided. For reference, model performance with all six features types is included in the last row.

None of the feature subsets improve over the all-feature performance, however, several combinations perform the same or similar to it. The two combinations with the smallest feature set (#1 and #2), employ only two out of the six feature types, and yet provide surprisingly strong performance: F1 of 0.80 and 0.81, respectively. The model trained with the union of these two feature combinations (#4. POS tag count + NE + Embeddings) performs on par with all-features model, suggesting large amount of feature redundancy (three out of six feature types: 50%). However, not any subset of three feature types is equivalent – #5 and #6 also employ three feature types but perform poorly as compared to the all-feature model. Overall, this study illustrated the importance of different feature types, and identified feature combination that reduced the number of features without degrading the model performance.

C. Attainment CCT: RF Model

The RF model trained with POS tag count, NE and Embeddings features scored F1-score of 0.81 for Attainment CCT classification. Table V shows the performance of the RF model. The model scored 0.88 recall for Class 1, same as LR model. However, it also mis-classified 16 non-Attainment sentences thus reducing Class 1 precision to 0.59. The RF model is seems to be biased toward the Class 1 similar to LR model. The majority of mis-classified Class 0 sentences were similar to Attainment CCT excerpts but with no explicit reference of career or education goals. This is due to higher feature weights assigned for word embeddings. When RF model was trained using all the feature sets discussed in III-B1 and III-B2, model F1 score reduced to 0.80 because of large number of feature set compared to training data.

D. Attainment CCT: LSTM Model

The LSTM model performed better than the LR and RF model due to its ability to understand contextual information of passed sequence. After 25 epoch of batch size 256, model scored F1-score of 0.84. Table VI shows the LSTM model results. The recall of Class 0 improved to 0.94 and precision of Class 1 improved to 0.76 by correctly classifying Class 0 sentences which were missed by LR and RF models. However, the number of true positives also reduced in LSTM network and thus reducing the recall of Class 1 to 0.73.

E. Attainment CCT: BERT Model

The BERT$_{BASE}$ model pre-trained on huge general domain corpus is fine tuned with training dataset of 1270 sentences. All the other Attainment classification model results are surpassed by the BERT$_{BASE}$ model which achieved F1-score of 0.92. Table VII shows the result for BERT fine tuned for Attainment CCT data. The BERT model correctly predicted 24 sentences with Attainment out of 26 and thus improving Class 1 recall to 0.92. With the vast contextual information

Model	Precision	Recall	F1
Logistic Regression	0.79	0.88	0.82
Random Forest	0.78	0.87	0.81
LSTM	0.85	0.84	0.84
BERT	**0.90**	**0.94**	**0.92**

TABLE VIII
COMPARISON OF MODEL PERFORMANCES - ATTAINMENT CCT CLASSIFICATION

	Precision	Recall	F1	DataPoints
Class 0	0.97	0.98	0.98	62
Class 1	0.89	0.80	0.84	10
Macro Average score	0.93	0.89	0.91	72

TABLE IX
RESULTS WITH LOGISTIC REGRESSION FOR FIRST GENERATION CCT CLASSIFICATION

	Precision	Recall	F1	DataPoints
Class 0	0.94	0.98	0.96	62
Class 1	0.86	0.60	0.71	10
Macro Average score	0.90	0.79	0.83	72

TABLE X
RESULTS WITH BERT FOR FIRST GENERATION CCT CLASSIFICATION

processed by the BERT pre-trained model, the false positives are also reduced and hence improving the precision of Class 1 to 0.83.

The high performance of BERT model can be explained based on the bi-directional self attention mechanism of the model pre-trained on a total of 3300 million words consisting of contextual information. Table VIII lists the result of all 4 models of the TACCTI system for Attainment CCT classification. As we can see, BERT model achieves the highest score for precision and recall and hence better F1-score.

F. First Generation Classification Models

The base model of First Generation CCT is built with Logistic Regression trained with unigrams as features. The Logistic Regression performed well with F1-score of 0.91. Table IX lists the Precision, Recall and F1 score of First Generation base model with only unigrams as a feature. This result is obtained on highly skewed dataset of 14% Class 1 and 86% Class 0 datapoints.

To check if model is over-fitting, the training dataset is used to perform stratified 9 fold cross validation. The average F1 score of 9 fold cross validation with Logistic Regression model and unigram features is obtained as 0.83. Looking at the difference between F1-score of the model on test data and average F1-score obtained during cross validation, model could be over-fitting due to small amount of training data.

When the same training data was used to train BERT$_{BASE}$ model for First Generation CCT classification, it achieved F1 score to 0.83. Because of the small training data, BERT$_{BASE}$ model did not perform well compared to Logistic Regression in case of First Generation CCT classification. Table X shows the result for BERT model for First Generation CCT classification. Out of 10 Class 1 sentences with First Generation CCT in it, BERT could correctly classify only 6 and thus scored less recall for Class 1. Table XI shows First Generation classification model comparison. Even though Logistic Regression model performed well with 0.91 F1 score, looking at the cross-validation results, model might have over-fitted with less training data.

Model	Precision	Recall	F1
Logistic Regression	**0.93**	**0.89**	**0.91**
BERT	0.90	0.79	0.83

TABLE XI
COMPARISON OF MODEL PERFORMANCES- FIRST GENERATION CLASSIFICATION

VI. DISCUSSION

The introduction of Named Entity (NE) and STEM similarity features improved the performance of the LR model for Attainment CCT. As the NE feature has more weight assigned during the training, the LR model mis-classifies few sentences. For example, "*my aunt is someone i look up to since she is a **radiologist** at the stanford medical center in palo alto*" is predicted as Class label 1 because of keyword *radiologist* which is a named entity of type profession. The LR model still lacks the ability to handle negation in the sentence. In the sentence, "*however i am **unsure** if i want to continue down the path of preparing for medical school and may decide to do something else with my major*", the LR model was unable to grasp the context of negation and predicted the sentence as Class label 1. Even though word embedding improves the model performance, sentences which looks like expressing Attainment CCT but with no explicit mention of career or education goals such as, "*i am taking physics because i will need it in order to pursue my major and i want to expand my knowledge in the field of math and sciences*" is classified as Class 1. The BERT model pre-trained on huge general corpus to include the contextual information, helps to correctly classify these datapoints and reduce the number of false positives. Few of the classification errors in Attainment CCT models are due to annotation discrepancies.

- The sentence "*i'm in stem to have my requirements met for medical school*" was missed during annotation, but all the models correctly classified this as Attainment CCT.
- The sentence "*i am here because i know i want to do well in this physics class*" does not really fit the thumb-rule for Attainment CCT since it doesn't contain any specific goal related to education or career. This is predicted as Class label 0 by the classification models but incorrectly labeled as Class label 1 during annotation.

We performed the random under-sampling on Attainment training data to check the performance of balanced dataset, which consisted 228 Class 1 and Class 0 sentences. The Logistic Regression model with NER, POS tag count and word embedding features was trained with balanced training data and tested on the original imbalanced test data to mimic real-world model usage. The F1 score of the model dropped from 0.82 to 0.81 as model mis-classified 18 non-attainment sentences lowering precision of Class 1 to 0.57.

Since the Logistic Regression model for First Generation CCT classification is trained on only unigrams features, it failed to detect the First Generation CCT when it is not explicitly mentioned using *First Generation phrases* like *first in my family* or *I am first generation*. For example,

- *It is unfortunate that my parents never went to college so they never really support me when I need because they dont know how to*
- *Neither of the both sides of my family had went to school or even had an easy life after high school*

Because of the lack of context information, Logistic Regression model also miss-classify the non-First Generation sentences as Class 1. In the case of *My parents are first generation immigrants so it was imperative that they have a support system when they migrated to the U(S)*, model classified the sentence as Class label 1 because of the presence of phrase *first generation* and model was unable to comprehend the context of the sentence mentioning parents.

Even though BERT model fine tuned on First Generation did not perform well due to lack of data, based on the learning ability of BERT, the results for First Generation CCT would definitely improve with the more training data as BERT model can capture the context of the sentence.

VII. Conclusion

The Alma project aims to help HU students in STEM by introducing reflective journals with prompts such as 'why am I here' to assert their life purpose and career motivation. The Alma research team has identified 11 different cultural capital themes (CCTs) in the student journals which reflect the skills, ambitions, challenges and life experience of HU students in STEM program. To assist with the annotation process, we have proposed TACCTI (pronounced as *Tacit*)), Text analysis and mAchine learning for Cultural Capital Theme - a scalable computational framework to identify instances of CCTs in the student essays.

In this work, we have focused on detecting presence (or absence) of two CCTs – Attainment and First Generation. Attainment could be explicit reference of the career or educational goals whereas First Generation CCT signifies student being first in their family to go to college. We found that using essays as datapoints for classification does not provide the granularity for assigning feature weights accurately. Hence essays are split into sentences and sentence-level data has been used for classification models in TACCTI. For Attainment CCT classification, models are built using both hand-crafted features – Logistic Regression and Random Forest and deep learning models – LSTM and BERT. For models using hand-crafted features, we found that, word embeddings and named entity type of features had higher effect on performance. Comparison of all four models indicates that, with the pre-trained contextual information, the BERT model achieves near-human precision for identifying Attainment CCT with an effectiveness of 0.92. For the First Generation CCT classification, with the limited available data, we built a Logistic Regression model with unigram features and a deep learning model using BERT. Upon analyzing we found that, Logistic Regression performed better than the BERT model when training data is limited. The Logistic Regression model for First Generation classification achieved effectiveness of 0.91 but faced the issue of overfitting. However, with more training data, the model performance can be improved further for First Generation CCT classification.

Acknowledgment

We thank Alegra Eroy-Reveles for important intellectual contributions including framing of the Alma Project, and Mireya Arreguin, Imani Davis, Amal Egeh, Anjelica Jones, Alex Macha-Lopez, Michaela Perez, and Rachel Xie for performing thematic analyses, and the many STEM students who have thoughtfully shared and participated in the reflective journaling. This project is supported in part by SF BUILD and SFSU College of Science and Engineering Dean's Office. SF BUILD is funded by the NIH Common Fund Linked grants: UL1 GM118985, TL4 GM118986, and RL5118984.

References

[1] E. Mcgee, B. Thakore, and S. LaBlance, "The burden of being"model": Racialized experiences of asian stem college students." Journal of Diversity in Higher Education, 2016.
[2] M. Estrada, A. Eroy-Reveles, and J. Matsui, "The influence of affirming kindness and community on broadening participation in stem career pathways." Social issues and policy review, 12 1:258–297, 2018.
[3] M. Estrada, M. Burnett, and AG. Campbell et al, "Improving underrepresented minority student persistence in stem." CBE-Life Sciences Education,15(3), 2016.
[4] T. Yosso," Whose culture has capital? a critical race theory discussion of community cultural wealth." Race Ethnicity and Education, 8(1):69–91, 2015.
[5] K. Tran, K. Coble, and A. Eroy-Reveles, "The Alma Project: Reflecting on Indigenous Knowledge in the 21st Century." 524:119, 2019.
[6] V. Kanagala, A. Nora, and L. Rendon, "A framework for understanding latino/a cultural wealth." Diversity Democracy, 19(1), 2016.
[7] A. Liaw and M. Wiener, "Classification and regression by randomforest." R News, 2(3):18–22, 2002.
[8] M. Schuster and K.K. Paliwal, "Bidirectional recurrent neural networks." IEEE Transactions on Signal Processing, vol. 45, no. 11, pp. 2673-2681, doi: 10.1109/78.650093, 1997.
[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: pre-training of deep bidirectional transformers for language under-standing." arXiv preprint arXiv:1810.04805, 2018.
[10] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit." pp. 55–60, 2014.
[11] F. Pedregosa, G. Varoquaux, A. Gramfort, and et al, "Scikit-learn: Machine learning in Python." Journal of Machine Learning Research,12:2825–2830, 2011.
[12] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation." in EMNLP, vol. 14, pp. 1532–1543, 2014.
[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space." ICLR Workshop, 2013.
[14] L.C. Moll, C. Amanti, D. Neff, and N. González, "Funds of Knowledge for Teaching: Using a Qualitative Approach to Connect Homes and Classrooms." Theory Into Practice ,31 (2): 132–41, 2001.
[15] A. Gloria, Borderlands: La Frontera: The New Mestiza, 2nd ed. San Francisco, CA: Aunt Lute Books, 1999.
[16] D.D. Bernal, "Learning and Living Pedagogies of the Home: The Mestizo Consciousness of Chicana Students." International Journal of Qualitative Studies in Education 14 (5): 623–39, 2010.
[17] T. Marwan, I. Mai and E.-M. Nagwa, "Imbalanced Toxic Comments Classification Using Data Augmentation and Deep Learning." 10.1109/ICMLA.2018.00141, 2018.
[18] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification?" CoRR, abs/1905.05583, 2019.
[19] G. Miller, "WordNet: a Lexical Database for English." Communications of the ACM, 38(11):39–41, 1995.
[20] S. Loria, TextBlob documentation, unpublished.
[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory." Neural Computation, 9(8):1735–1780, 1997.