Enhancing STEM Motivation through Personal and Communal Values: NLP for Assessment of Utility Value in Student Writing

Beata Beigman Klebanov[1] **Jill Burstein**[1] **Judith M. Harackiewicz**[2]
Stacy J. Priniski[2] **Matthew Mulholland**[1]
[1]Educational Testing Service
[2]University of Wisconsin, Madison
bbeigmanklebanov,jburstein,mmulholland@ets.org
jmharack,spriniski@wisc.edu

Abstract

We present, to our knowledge, the first experiments on using NLP to measure the extent to which a writing sample expresses the writer's *utility value* from studying a STEM subject. Studies in social psychology have shown that a writing intervention where a STEM student is asked to reflect on the value of the STEM subject in their personal and social life is effective for improving motivation and retention of students in STEM in college. Automated assessment of UV in student writing would allow scaling the intervention up, opening access to its benefits to multitudes of college students. Our results on biology data suggest that expression of utility value can be measured with reasonable accuracy using automated means, especially in personal essays.

1 Introduction

Motivational factors, such as goals, confidence, interest and values have been shown to be important in supporting continuing engagement and success in academic pursuits at all age levels (Pintrich, 2003).

In recent years a number of promising interventions have been developed in the field of empirical social psychology to promote student motivation. Among the most successful of these interventions in college classes is the Utility Value Intervention (**UVI**) (Harackiewicz et al., 2014; Harackiewicz et al., 2015). Grounded in Eccles' Expectancy-Value Theory (Eccles et al., 1983; Eccles, 2009), the UVI, in which students write about the personal relevance of course material, helps students discover connections between course topics and their lives – in their own terms. Discovering these connections helps students appreciate the value of their course work, leading to a deeper level of engagement with course topics that, in turn, improves performance. The effectiveness of these UVI writing assignments has been demonstrated with experimental laboratory studies and field experiments in college and high school (Canning and Harackiewicz, 2015; Harackiewicz et al., 2015; Gaspard et al., 2015; Hulleman et al., 2010; Hulleman and Harackiewicz, 2009). These UVIs are most effective for promoting motivation among those most at risk for dropping out (Harackiewicz et al., 2015; Hulleman and Harackiewicz, 2009; Hulleman et al., 2010).

A large-scale application of UVI in college and other school contexts is hindered by the need to train and employ humans to score students' writing samples for utility value. Our goal is to assess the potential of NLP to provide an automated UV evaluation that could, in turn, support scaling up the UVIs to reach many more struggling college freshmen. An automatically delivered and scored UVI would allow STEM faculty to assign UVI as homework; the automatic scores would be delivered to faculty, and students whose writing samples had insufficient expression of utility would be routed to a one-on-one session with the instructor or a teaching assistant, to discuss their plans and values to help find personal utility in studying STEM.

2 Data

Materials used in our experiments come from the study by Harackiewicz et al. (2015).[1] They col-

[1]For data contact Prof. Harackiewicz, jmharack@wisc.edu.

Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pages 199–205,
San Diego, California, June 16, 2016. ©2016 Association for Computational Linguistics

lected writing samples from first-year students enrolled in introductory biology courses at University of Wisconsin, Madison, 2012-2014. Students were asked to pose a question related to the recently studied module and answer it while incorporating utility value (**UV**), that is, explaining how the biology topic was related to their own or other people's lives. Six different biology topics are covered in the dataset (e.g., cell biology, ecology).

The utility value and control writing assignments were coded by research assistants for the level of utility value articulated in each essay, on a scale of 0-4, based on how specific and personal the utility value connection was to the individual. A "0" on this scale indicates no utility; a "1" indicates general utility applied to humans generically; a "2" indicates utility that is general enough to apply to anyone, but is applied to the individual; a "3" indicates utility that is specific to the individual; and a "4" indicates a strong, specific connection to the individual that includes a deeper appreciation or future application of the material. Inter-rater reliability with this coding rubric was high, with two independent coders providing the same score on 91% of essays. Disagreements were resolved by discussion.

Students were given 5 days to complete the assignment. Each student contributed 3 writing samples, in same or different genres, as described below.

Genre Variation

Students were assigned one of the following four genres, or given a choice (usually between Essay and Letter). The Essay, Letter, and Society genres are UVI genres, in that they request reference to utility value, whereas Summary is a control genre that only asks for a summary of the course material.

Assignment (common to all genres): Select a concept or issue that was covered in lecture and formulate a question.

Letter Write a 1-2 page letter to a family member or close friend, addressing this question and discuss the relevance of this specific concept or issue to this other person. Be sure to include some concrete information that was covered in this unit, explaining why the information is relevant to this person's life, or useful for this person. Be sure to explain how the information applies to this person and give examples.

Essay Write an essay addressing this question and discuss the relevance of the concept or issue to your own life. Be sure to include some concrete information that was covered in this unit, explaining why this specific information is relevant to your life or useful for you. Be sure to explain how the information applies to you personally and give examples.

Society Write an essay addressing this question and discuss the relevance of the concept or issue to people or society. Be sure to include some concrete information that was covered in this unit, explaining why this specific information is relevant to people's lives and/or useful for society and how the information applies to humans. Be sure to give examples.

Summary Select the relevant information from class notes and the textbook, and write a 1-2 page response to your question. You should attempt to organize the material in a meaningful way, rather than simply listing the main facts or research findings. Remember to summarize the material in your own words. You do not need to provide citations.

To exemplify UV-rich writing, consider the following excerpt from a Letter on Ecology:

> I heard that you are coming back to America after retirement and are planning on starting a winery. I am offering my help in choosing where to live that would promote the growth of grapes the best. Grapes are best grown in climates that receive large amounts of sunlight during the growing season, get moderate to low amounts of water, and have relatively warm summers. I highly recommend that you move to the west coast, and specifically the middle of the coast in California, to maximize the efficiency of your winery. **Letter, Ecology**

Table 1 shows data partition sizes and average essay length per genre. We note that the test set contains writing samples from unseen students.[2] Table 2 shows the UV score distributions in the training data.

[2]not unseen essays from students who contributed another writing sample to the train set

Genre	Number of Samples			Av. Length
	TRAIN	DEV	TEST	(words)
Essay	2,766	840	329	508
Letter	2,457	867	266	508
Society	273	84	44	492
Summary	3,353	1,160	345	486

Table 1: Summary of data, by genre

Genre	UV Score				
	0	1	2	3	4
Essay	.04	.15	.09	.38	.34
Letter	.02	.03	.04	.32	.59
Society	.03	.75	.02	.16	.04
Summary	.59	.38	.00	.02	.01

Table 2: Distributions of utility value score, by genre

3 Features

For measuring utility value in a writing sample, we developed a set of features that address the form and the content of personalized writing.

3.1 Pronouns

We expect grammatical categories that signal reference to self, addressee, or other humans to occur frequently in UV-rich writing. We calculate log frequency per 1,000 words for the following categories:

- PRO_SG1: First person singular pronouns
- PRO_PL1: First person plural pronouns
- PRO_2: Second person pronouns
- DET_POS: Possessive determiners (e.g., their)
- PRO_INDEF: Indefinite pronouns (e.g., anyone)

3.2 General Vocabulary

Since expression of UV is likely to refer to everyday concerns and activities, we expect essays rich in UV to be less technical, on average, than essays that only summarize the technical content of a biology course, and therefore use shorter, more common, and more concrete words, as well as a larger variety of words. We define the following:

- WORDLN: Average word length (in letters)
- WF_MEDIAN: Median word frequency
- ACADEMICWL: Proportion of academic words (Coxhead, 2000) in content words in the essay
- CONCRETE: Log frequency per 1,000 words of words from the MRC concreteness database (Coltheart, 1981)
- TYPES: # of different words (types count)

3.3 Genre-Topic Vocabulary

We define a feature that captures use of language that is common for the given genre in the given topic, under the assumption that, for example, different personal essays on ecology might pick similar subtopics in ecology and also possibly present similar UV statements. For a given writing sample in genre G on topic T, we identify words that are typical of the genre G for the topic T (genre-topic words). A word is typical of genre G for the topic T if it occurs more frequently in genre G on topic T than in all other genres taken together on topic T.[3] The estimation of typical genre-topic vocabulary is done on training and development data.

- GENREVOC: Log type proportion of genre-topic words.

3.4 Argumentative and Narrative Elements

While summaries of technical biology material are likely to be written in an expository, informational style, one might expect the UV elements to be more argumentative, as the writer needs to put forward a claim regarding the relationship between their own or other people's lives and biology knowledge, along with necessary qualifications. We therefore defined lists of expressions that could serve to develop an argument (based on Burstein et al. (1998)) and a list of expressions that qualify or enhance a claim (based on Aull and Lancaster (2014)). The features use log token count for each category.

- ARGDEV: Words that could serve to develop an argument, such as *plausibly*, *just as*, *not enough*, *specifically*, *for instance*, *unfortunately*, *doubtless*, *for sure*, *supposing*, *what if*.

[3]This is similar to Lin and Hovy (2000) topic signatures approach (or, rather, genre-topic signatures here), without the transformation that supports significance thresholds. This simpler approach was found to be effective in our work on topicality for essay scoring (Beigman Klebanov et al., 2016).

- HEDGEBOOST: Hedging and boosting expressions, such as: *perhaps*, *probably*, *to some extent*, *not entirely true*, *less likely*, *roughly* (hedges); *naturally*, *can never*, *inevitably*, *only way*, *vital that* (boosters).

In addition, in order to connect the biology content to the writer's own life, the writer might need to provide a personal mini-narrative – background with details about the events in his or her life that motivate the particular UV statement. Since heavier reliance on verbs is a hallmark of narrativity, we define the following features (using log frequency per 1,000 words):

- COMVERBS: Common verbs (*get*, *go*, *know*, *put*, *think*, *want*)
- PASTTENSEVERBS: VBD part-of-speech tags

3.5 Likely UV content

Building on our observations of common UV content in the training data and on previous work by Harackiewicz et al. (2015), we capture specific content and attitude using dictionaries from LIWC (Pennebaker et al., 2007). In particular, UV statements often mention the benefit of scientific knowledge for improving understanding and for avoiding unnecessary harm and risk; specific themes often include considerations of health and diet. For each category, we use log proportion of words belonging to the category in the given writing sample as a feature.

- AFFECT: Words expressing positive and negative affect, such as *love*, *nice*, *sweet* and *hurt*, *ugly*, *nasty*, respectively.
- SOCIAL PROCESSES: Words expressing social relations and interactions, such as *talk*, *mate*, *share*, *child*, as well as words in the LIWC categories of *Family*, *Friends*, and *Humans*.
- INSIGHT: Words that signify cognitive engagement, such as *think*, *know*, *consider*.
- HEALTH: Words that refer to matters of health and disease, such as *clinic*, *flu*, *pill*.
- RISK: Dangers and things to avoid.
- INGESTION: Example words: *eat*, *dish*, *pizza*.

4 Evaluation

4.1 Feature Families

We evaluated each of the five feature families on its own for predicting the UV score, as well as the added value of each feature family over the combination of all the other families. On the development set, we experimented with a number of machine learning algorithms using scikit-learn toolkit (Pedregosa et al., 2011) via SKLL:[4] random forest regressor, elastic net regressor, linear regression, linear support vector regression, ridge regression, support vector regression with RBF kernel. Random forest was selected as it showed the best average performance across the four genres. Pearson correlation (r) is the objective function. Table 3 presents the results on the development set: The first 5 rows show each feature family on its own, followed by ALL (all families together) and by models where one family was ablated at a time.

Feature Family	Essay	Letter	Soc.	Sum.
Pronouns	.759	.442	.527	.544
General Voc	.302	.200	*.165*	.260
GenreVoc	.219	.378	*.186*	.377
ArgNarr	.289	.286	.249	.195
UV content	.306	.313	*.025*	.318
ALL	.784	.543	.527	.622
– Pronouns	.451	.450	.309	.450
– General Voc	.777	.536	.527	.611
– GenreVoc	.787	.500	.527	.586
– ArgNarr	.774	.529	.527	.622
– UV content	.768	.542	.527	.622

Table 3: Pearson correlations with UV score for various feature families. Italicized correlations are not significant ($p > 0.05$).

We observe that all feature families attained statistically significant correlations with utility value scores in Essay, Letter, and Summary genres. With a single exception (GenreVoc in Essay), the correlation attained by the feature set that contains all the families (ALL) was the same or higher than in cases where a feature family was ablated. For Society, all features apart from Pronouns are quite weak, though not detrimental to performance. We decided to keep all features for the benchmark evaluation.

[4] https://github.com/EducationalTestingService/skll, version 1.1.1.

4.2 Benchmark Evaluations

We compare the effectiveness of our feature set designed specifically for this task with word ngrams (n=1,2,3) baseline. Using the development data, we evaluated ngrams using the same set of machine learning algorithms as for the experimental features (see section 4.1), and selected elastic net regression due to best average performance across the genres. Table 4 shows the performance of the experimental system and the baseline, on the blind test set, per genre. The experimental feature set (containing the total of 21 features) is on par with the baseline, on average, while showing worse performance on Letters, and better performance on Society, the dataset with the smallest number of instances.

Next, we combined the baseline and the experimental feature sets, using the elastic net regressor. Table 4 row Baseline+Exp shows the performance of the combined feature set on test data. The combination yields an average relative improvement of 4% over the baseline and 5.7% – over the experimental system. Generally, the combination matches the better performance between baseline and experimental features on Essay, Summary, and Society genres, and improves over the best performance in the most difficult Letters genre (6.8% relative improvement over baseline).

System	Essay	Letter	Soc.	Sum.	Av.
Baseline	.788	.437	.731	.662	.655
Experimental	.786	.358	.799	.633	.644
Baseline+Exp	.798	.467	.796	.663	.681

Table 4: Evaluation vs ngrams baseline on test data.

5 Related Work

Harackiewicz et al. (2015) performed an exploratory analysis of UV writing versus control (Summary) writing, using a subset of categories from LIWC. They found that the categories of personal pronouns, words referencing family, friends, and other humans, as well as words describing social processes were used in significantly higher proportions in UV writing. They also found that words of cognitive involvement and insight are used more in UV writing.

In recent years, NLP techniques have increasingly been applied to studying a variety of social and psychological phenomena. In particular, NLP research has been used to detect language that reflects certain traits of the authors' disposition or thinking, such as detection of deception, sentiment and affect, flirtation, ideological orientation, depression, and suicidal tendencies (Mihalcea and Strapparava, 2009; Abouelenien et al., 2014; Hu and Liu, 2004; Ranganath et al., 2009; Neviarouskaya et al., 2010; Beigman Klebanov et al., 2010; Greene and Resnik, 2009; Pedersen, 2015; Resnik et al., 2013; Mulholland and Quinn, 2013). Such studies have a tremendous potential to help measure, understand, and ultimately enhance personal and societal well being. We believe that the line of inquiry initiated by the current study that is focused on motivation in college likewise promises important potential benefits.

6 Conclusion & Future Work

We presented the first experiments, to our knowledge, on using NLP to measure the extent to which a writing sample contains expression of the writer's utility value from studying a STEM subject. Studies in social psychology have shown that a writing intervention where a STEM student is asked to reflect on the value of the STEM subject in their personal and social lives is effective for improving motivation and retention of students in STEM subjects in college. However, the need for a trained human reader to score the writing samples has so far hindered an application of the intervention on a large scale. Our results on biology data are encouraging, suggesting that utility value can be measured with reasonable accuracy using automated means, especially in personal essays.

A direction of future work that would enable further progress in scaling up the UV writing intervention involves development of a support system for scaffolding the process of writing about UV. In particular, once the automated measurement has determined that a draft of an essay lacks sufficient expression of UV, more specific feedback and dedicated writing activities could be automatically suggested to facilitate the student's thinking and writing about the utility value of the STEM subject, which would help boost the student's motivation and success in STEM education.

7 Acknowledgement

We would like to thank Su-Youn Yoon and Diane Napolitano for their help with feature generation; Keelan Evanini, Jesse Sparks and Michael Flor for their comments on an earlier draft of this paper.

References

M. Abouelenien, V. Perez-Rosas, R. Mihalcea, and M. Burzo. 2014. Deception detection using a multimodal approach. In *Proceedings of the 16th ACM International Conference on Multimodal Interaction*, pages 58–65, New York. ACM.

Laura L. Aull and Zak Lancaster. 2014. Linguistic markers of stance in early and advanced academic writing: A corpus-based comparison. *Written Communication*, 31:151–183.

Beata Beigman Klebanov, Eyal Beigman, and Daniel Diermeier. 2010. Vocabulary choice as an indicator of perspective. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 253–257, Uppsala, Sweden, July. Association for Computational Linguistics.

Beata Beigman Klebanov, Michael Flor, and Binod Gyawali. 2016. Topicality-based indices for essay scoring. San Diego, CA, June. Association for Computational Linguistics.

Jill Burstein, Karen Kukich, Susanne Wolff, Ji Lu, and Martin Chodorow. 1998. Enriching automated essay scoring using discourse marking. In *Proceedings of the ACL Workshop on Discourse Relations and Discourse Marking*, pages 15–21, Montréal, Canada, August. Association for Computational Linguistics.

E. Canning and J. Harackiewicz. 2015. Teach it, don't preach it: The differential effects of directly communicated and self-generated utility-value information. *Motivation Science*, 1:47–71.

M. Coltheart. 1981. The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, 33A:497–505.

Averil Coxhead. 2000. A new academic word list. *TESOL Quarterly*, 34(2):213–238.

J. Eccles, T. Adler, R. Futterman, S. Goff, Kaczala C., and J. Meece. 1983. Expectations, values and academic behaviors. In J. T. Spence, editor, *Perspective on achievement and achievement motivation*, pages 75–146. San Francisco, CA: W. H. Freeman.

J. Eccles. 2009. Who am I and what am I going to do with my life? Personal and collective identities as motivators of action. *Educational Psychologist*, 44:78–89.

H. Gaspard, A. Dicke, Flunger B., M. Brisson, I. Hafner, B. Nagengast, and U. Trautwein. 2015. Fostering adolescents' value beliefs for mathematics with a relevance intervention in the classroom. *Developmental Psychology*, 51:1226–1240.

Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 503–511, Boulder, Colorado, June. Association for Computational Linguistics.

J. Harackiewicz, Y. Tibbetts, E. Canning, and J. Hyde. 2014. Harnessing values to promote motivation in education. volume 18, pages 71–105. Bingley, UK: Emerald Group Publishing Limited.

J. Harackiewicz, E. Canning, Y. Tibbetts, S. Priniski, and J. Hyde. 2015. Closing achievement gaps with a utility-value intervention: Disentangling race and social class. *Journal of Personality and Social Psychology, http:// dx.doi.org/10.1037/pspp0000075*.

M. Hu and B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177, Seattle, Washington. ACM.

C. Hulleman and J. Harackiewicz. 2009. Promoting interest and performance in high school science classes. *Science*, 326:1410–1412.

C. Hulleman, O. Godes, B. Hendricks, and J. Harackiewicz. 2010. Enhancing interest and performance with a utility value intervention. *Journal of Educational Psychology*, 102:880–895.

C. Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of COLING*, pages 495–501.

R. Mihalcea and C. Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*, pages 309–312, Singapore. Association for Computational Linguistics.

Matthew Mulholland and Joanne Quinn. 2013. Suicidal tendencies: The automatic classification of suicidal and non-suicidal lyricists using nlp. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 680–684, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2010. Recognition of affect, judgment, and appreciation in text. In *Proceedings of the 23rd International Conference on Computational Linguistics*

(Coling 2010), pages 806–814, Beijing, China, August. Coling 2010 Organizing Committee.

T. Pedersen. 2015. Screening twitter users for depression and ptsd with lexical decision lists. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 46–53, Denver, Colorado. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

J. Pennebaker, R. Booth, and M. Francis. 2007. *Linguistic Inquiry and Word Count: LIWC (Computer software)*. Austin, TX: LIWC.net.

P. Pintrich. 2003. A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology*.

Rajesh Ranganath, Dan Jurafsky, and Dan McFarland. 2009. It's not you, it's me: Detecting flirting and its misperception in speed-dates. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 334–342, Singapore, August. Association for Computational Linguistics.

Philip Resnik, Anderson Garron, and Rebecca Resnik. 2013. Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1348–1353, Seattle, Washington, USA, October. Association for Computational Linguistics.