Int J Artif Intell Educ (2017) 27:791–818
DOI 10.1007/s40593-017-0141-4

ARTICLE

Reflective Writing About the Utility Value of Science as a Tool for Increasing STEM Motivation and Retention – Can AI Help Scale Up?

Beata Beigman Klebanov[1] · **Jill Burstein**[1] · **Judith M. Harackiewicz**[2] · **Stacy J. Priniski**[2] · **Matthew Mulholland**[1]

Published online: 9 March 2017

Abstract The integration of subject matter learning with reading and writing skills takes place in multiple ways. Students learn to read, interpret, and write texts in the discipline-relevant genres. However, writing can be used not only for the purposes of practice in professional communication, but also as an opportunity to reflect on the learned material. In this paper, we address a writing intervention – Utility Value (UV) intervention – that has been shown to be effective for promoting interest and retention in STEM subjects in laboratory studies and field experiments. We conduct a detailed investigation into the potential of natural language processing technology to support evaluation of such writing at scale: We devise a set of features that characterize UV writing across different genres, present common themes, and evaluate UV scoring models using essays on known and new biology topics. The automated UV scoring results are, we believe, promising, especially for the personal essay genre.

✉ Beata Beigman Klebanov
bbeigmanklebanov@ets.org

Jill Burstein
jburstein@ets.org

Judith M. Harackiewicz
jmharack@wisc.edu

Stacy J. Priniski
spriniski@wisc.edu

Matthew Mulholland
mmulholland@ets.org

[1] Educational Testing Service, Princeton, NJ, 08541, USA

[2] University of Wisconsin, Madison, WI, 53706, USA

Keywords Intrapersonal factors · Motivation · Automated writing evaluation · Utility value · Natural language processing · Machine learning

Introduction

Approximately fifty percent of entering college students intending to major in **STEM** (Science, Technology, Engineering, and Mathematics) leave to pursue other majors or drop out of college altogether (NCES 2013, 2014). This attrition rate is alarming because projections indicate a shortfall of approximately one million STEM professionals in the next decade (PCAST 2012). Three-quarters of that projected shortfall could be addressed by retaining an additional ten percent of the students who leave.

Keeping students interested in science courses is crucial to retaining them in STEM majors and on track for STEM careers. One way to develop interest in activities is to find meaning and value in those activities (Durik and Harackiewicz 2007; Hidi and Harackiewicz 2000), and one type of task value that has proven to be a powerful predictor of interest, effort, and performance is utility value (**UV**). A person finds UV in a task if he or she believes it is useful and relevant beyond the immediate situation, for other tasks or aspects of a person's life. For example, "I will really need this for medical school," or "This material will be important when I take over the family farm."

Correlational research indicates that when students perceive value in course topics, they work harder, develop more interest, and perform better (Harackiewicz et al. 2008; Hulleman et al. 2008; Wigfield 1994). They are also more likely to take additional courses and complete their degree programs (Harackiewicz et al. 2008; Hulleman et al. 2008). Students who see the utility value of a field of study experience greater involvement and more positive task attitudes, and feel more identified with the domain (Brown et al. 2015; Smith et al. 2015). Thus the perception of UV can promote a student's sense of identity in a domain (Eccles 2009).

Recent experimental research suggests that it is possible to promote perceived UV with simple interventions that ask students to write about the relevance of course topics to their own life. These writing interventions in which students write essays connecting specific course content to their own lives work best for students who doubt their competence and have a history of poor performance. For example, Hulleman and Harackiewicz (2009) found that their Utility Value Intervention (**UVI**) raised interest and grades for 9th grade science students who had low performance expectations, relative to a control group. Hulleman et al. (2010) found that the same type of UVI promoted interest in an introductory psychology class for students who had performed poorly on early exams, relative to a control group. Recently, in a double-blind randomized field experiment conducted in an introductory college biology course, Harackiewicz et al. (2016) found that the UVI boosted course performance for all students, but was particularly effective among those who are traditionally most underrepresented in STEM (i.e., underrepresented minority students and first-generation college students). In summary, the UVI has been shown to increase motivation and performance in high school and college students.

The UVI can be integrated into course pedagogy, and has proven to be particularly beneficial for students with low success expectancies and/or with a record of low performance in a course. Implementation of the UVI is also consistent with pedagogical goals for science courses, where there has been a greater emphasis on writing in scientific domains (Brewer and Smith 2011). Indeed, writing assignments such as the UVI can help students learn to generate, justify and evaluate scientific arguments (Prain and Hand 2016). Writing activities embedded in science courses have been shown to enhance learning in biology and other domains (Gunel et al. 2007; Prain and Hand 2016).

Current UVIs use human evaluations of expressions of UV in an essay by specially trained research assistants; the procedure is thus time-intensive and costly. In this paper, we evaluate the extent to which natural language processing (NLP) technology can supplement or replace these human evaluations, by identifying linguistic markers related to reflective writing and combining them using machine learning to provide an overall assessment of the utility value expressed in a student's written piece. If successful, automated evaluation of utility expression could help scale UVI interventions up beyond research studies. This would allow disciplinary instructors, for example, to assign UVI to students as homework to be performed using a technological platform; the automatically generated utility value score would then be reported to the instructor. The large majority of students would complete the UVI assignment successfully; students who receive a low UV score apparently had difficulty articulating UV and might need additional help in seeing the relevance of course material to their personal and social lives. Such targeted assistance to specific students can be delivered on a one-to-one basis by the instructor or a teaching assistant. We envision that with the maturation of the NLP technology for analysis and assessment of reflective writing of this kind it would be possible to provide scaffolding while the students are working on the UVI assignment, in the form of pre-writing and interim writing tasks to help students learn to make and clearly articulate UV connections to course content.

This paper is organized as follows. First, we present the UTILITY VALUE INTERVENTION and the DATA used for the current study. We follow with a QUALITATIVE ANALYSIS OF UTILITY VALUE IN STUDENT WRITING, to inspect what it is that students say when they are making connections between personal and social life and STEM materials. We then present the FEATURES, namely, automatically extracted indicators, that will be used to build models for predicting the human-assigned utility value score of a given writing sample. A PRELIMINARY STUDY is then reported that evaluates the individual features and various combinations thereof in terms of correlations with human-assigned utility value scores. With the most promising features and feature combinations identified in the preliminary study using development data, we then conduct EXPERIMENT 1 where the chosen models are evaluated on samples written by new students, unseen during system training; system performance and some of its errors are discussed in that section. We then present EXPERIMENT 2, where we evaluate the ability of the computational models to predict utility value scores for samples written for biology topics that were not seen during system development. This evaluation sheds light on the extent to which the system can generalize to a new content area (though still within the discipline of biology). We follow with a review of RELATED WORK, DISCUSSION, and CONCLUSION.

Utility Value Intervention (UVI)

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). Competence and skills are necessary but not sufficient for academic success; promoting student interest and motivation is key to improving learning and persistence (Hidi and Harackiewicz 2000).

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, 2016). 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 (Canning and Harackiewicz 2015; Hulleman et al. 2010) and field experiments in college and high school biology courses (Harackiewicz et al. 2016; Hulleman and Harackiewicz 2009), introductory psychology courses in college (Hulleman et al. 2010), and high school math courses (Gaspard et al. 2015).

The materials used in our experiments come from the study by Harackiewicz et al. (2016). They collected writing samples from first-year students enrolled in introductory biology courses at University of Wisconsin, Madison, 2012-2014. Students were asked to write an essay or a letter posing a question related to the recently studied module and answering it while making sure to incorporate utility value (henceforth, **UV**), that is, how the biology topic was related to their own life (Essay) or the addressee's (Letter). 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. According to Harackiewicz et al. (2016), inter-rater reliability with this coding rubric was high ($\kappa = 0.88$), 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 are assigned one of the following four genres, or they are 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 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 excerpts from two writing samples, a Letter on Ecology and an Essay on Evolution:

> 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**.

> An example of a trait that is acquired but not heritable is fitness. I am an athlete, so I exercise regularly for my sport. However, fitness is a trait I have acquired in my lifetime, not one that was written in my genes at birth. This means that it is not heritable, so I cannot pass it on to my children. If I want my kids to participate in sports, I will have to encourage them to exercise and play sports so that they can acquire fitness. **Essay, Evolution**.

Method

In this study, we apply supervised machine learning methodology to evaluate the extent to which a writing sample can be automatically scored for expression of utility value. We use random forest regression (Breiman 2001)[1] to build prediction models using linguistic features, namely, automatically extracted indicators of various linguistic properties of the writing samples. All models are trained using Pearson's correlation (r) with human-assigned utility value scores as the objective function.

The data for every genre is partitioned into TRAIN, DEV (or development), and TEST sets.[2] The TRAIN set is used to automatically set the weights for the different features in a model that combines multiple features. Since we plan to evaluate multiple models that use different feature combinations, we use the DEV set for preliminary evaluations of a large number of models. Conceptually, this set is used to develop (or choose) good scoring models.[3] However, using only results on DEV set poses the risk of observing significant correlations by chance due to the large number of evaluations on the same data; the other hazard is the possibility of inadvertently over-fitting the data used for testing because various choices made on the basis of initial evaluations might end up being specially targeted to boost performance on that specific testing set. Therefore, a small number of best performing models from the preliminary evaluation are subsequently evaluated on the TEST set that contains fresh data, unseen during model training and development.

The TEST data is sampled by student, namely, if a student was selected to appear in TEST data, all three of her writing samples appeared in TEST sets for the relevant genres. The partition into TRAIN and DEV data is done randomly by essay, so it is possible for one writing sample of a student to appear in DEV data and two others – in TRAIN data.

In experiment 2, we use cross-validation methodology on the TRAIN data to evaluate the extent to which prediction models trained on writing samples addressing a given set of topics would generalize to a writing sample addressing a new topic. To this end, the TRAIN data is partitioned into six folds (subsets), each containing writing samples on one of the six topics represented in the dataset. For each round of cross-validation, we build models trained on five folds, and test on the sixth. We report results for each test fold, as well as the average across six evaluations, for every genre.

Data

For experiments reported in this paper, data were partitioned into TRAIN, DEV, and TEST sets, for each of the four genres. Table 1 shows the number of writing samples

[1] We used the implementation of random forest regressor in scikit-learn (Pedregosa et al. 2011) via SKLL (https://github.com/EducationalTestingService/skll) version 1.1.1

[2] See, for example, Ripley (1996) for a discussion of reasons for using training, validation (development), and testing sets in a typical experimental paradigm in supervised machine learning.

[3] The other common use of DEV set is to tune parameters for some of the features. We practice this use to a very limited extent, only in the construction of the GenreVoc feature, as will be explained in due course.

Table 1 Summary of data, by genre

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

in each of the sets, by genre. It is readily apparent that while Essay, Letter, and Summary genres have substantial amounts of data, the Society genre was administered relatively infrequently, and thus has relatively few writing samples. The average length of a writing sample across all genres is 499 words, standard deviation is 126 words.

The same data is also subdivided into six topics, corresponding to six modules in the biology course. The topical breakdown is shown in Table 2; the data are approximately balanced across topics.

Table 3 shows the distribution of human-assigned Utility Value scores by genre, in TRAIN data. The Summary genre is the control genre, where no UV was required; accordingly, 59% of the samples in this genre have the UV score of 0, and an additional 38% have the UV score of 1. In contrast, the Letter genre has 91% of its samples in UV scores of 3 or 4. The distribution in the Essay genre is more balanced, with 72% in the top two categories. The distribution in Society genre is also more spread out; we keep in mind, however, that there are relatively few essays available in this genre.

Qualitative Analysis of Utility Value in Student Writing

The main difference in the instructions for the different UVI genres (Essay, Letter, Society) is the way students are required to frame the utility value: value for oneself, value for another individual, and value for society, respectively. We therefore believe

Table 2 Summary of Data, by Topic

Topic	Number of Samples		Av. Length
	TRAIN + DEV	TEST	(words)
Evolution	1,570	144	496
Genetics	2,375	173	507
Cell biology	1,667	173	491
Animal Physiology	2,034	184	499
Plant Physiology	2,199	155	497
Ecology	1,957	155	505

Table 3 Distribution of utility value scores, by genre, in TRAIN data

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

that identifying words that characterize each of the UVI genres should help us understand how utility is expressed. Characterization of the Summary genre, in contrast, should tell us about language characteristics of non-UV texts since this is the control genre. In this section, we will investigate what vocabulary is pertinent to UV writing.

Illustration of Genre and Topic Influences on Vocabulary Distribution

To illustrate how word frequency can meaningfully differ across genres and topics, Fig. 1 shows frequencies per 10,000 words of the bigram "your body", across genres (in series) and across topics (x-axis). Clearly, "your body" is much more frequent in Letters than in any other genre, across topics (the blue line is always on top); however, "your body" is more frequent in Cell Biology and Animal Physiology topics than in the rest of the topics, for every single genre. So, this bigram is (1) genre-specific for Letters; (2) topic-specific for Cell Biology and Animal Physiology; and (3) task-

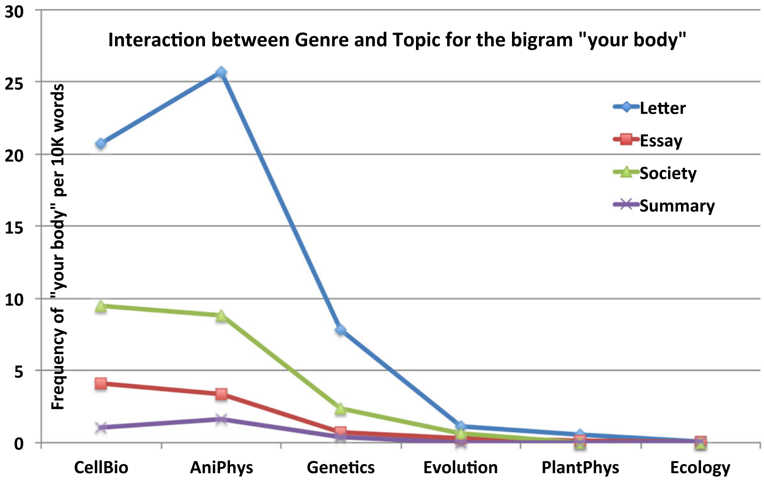

Fig. 1 Illustration of genre and topic dependence of vocabulary distribution

appropriate for referencing utility value, since it is more frequent in each of Letter, Essay, and Society genres than in the control Summary genre, across topics.

Inspecting Utility-Value Vocabulary

To find words that occur commonly in a given genre for a given topic, we use a frequency-based method. For each topic, we find words that have a higher frequency in the given genre (“in relevant documents”) than in other genres taken together (“in irrelevant documents”), for that topic. This provides a collection of words for Essay_Ecology that are more frequent there than in Letter_Ecology, Summary_Ecology, and Society_Ecology, taken together. This gives us a list of candidate genre_topic words, for each genre and topic combination; we call these **genre_topic sets**. Note that the same word can appear in multiple genre_topic sets.

To gain some insight into the vocabulary supporting the expression of utility value, we calculated a Word Utility Score (**WUS**) for each word as follows:

$$WUS(w) = -|\{t \in T | w \in Summary\_t\}| + \Sigma_{g\in\{Essay,Letter,Society\}} |\{t \in T | w \in g\_t\}| \tag{1}$$

where T is the set of topics T = {Cell Biology, Ecology, Evolution, Animal Physiology, Plant Physiology, Genetics}, g_t is the genre_topic set for genre *g* and topic *t*, and Summary_t is the genre_topic set for the Summary genre for topic *t*. A word with a high WUS would appear in genre_topic sets for multiple topics for the UVI genres (Essay, Letter, Society), and would tend to not appear in Summary genre_topic sets for the different topics. The maximal possible WUS is 18, corresponding to appearance in each of 3 UVI genres X 6 topics, and no appearances in any of the 6 Summary_topic sets. The two words with the highest WUS are *we* and *our*; they both received a score of 16, appearing in the genre_topic sets for all Essay and Letter topics, and in 4 out of 6 Society topics, but in none of the Summary genre_topic sets.

Table 4 shows some of the word families manually identified by the lead author in the 400 words with the highest WUS. The table reveals that UV-rich writing has a higher proportion of reference to people through 1st and 3rd person pronouns, generic terms (such as *people* or *person*), and reference to family and friends. This latter group can be partially explained by the request to write a letter to “a family member or close friend” in the instructions for the Letter genre. However, the words *children* and *friends* have a very high WUS of 14, as they are part of the genre_topic set not only in all Letter topics, but in all Essay topics and some Society topics as well – and in none of the Summary topics. Thus, even when people are addressing the instruction to write about utility to themselves, as in the Essay genre, they tend to reference other people, notably, children and friends.

The next four groups of words address the main roles of science in daily life – enabling improved understanding and knowledge, satisfying people’s curiosity, providing support and guidance in decision making, mitigating damage and addressing

Table 4 Word families identified in 400 words with the highest Word Utility Score (WUS)

Family Descriptor	High-WUS Words in the Family
1st Person	we our us ourselves I am me mine
3rd Person Pronoun	she her them him his he
Generic references to people	someone who everyone people anyone person girl men man fellow human teenager
Family & Friends	children child friends home family cousins father kids babies baby kid
Understanding, Reasoning, Curiosity	knowledge aware wonder understand how find knowing importance interest understanding knows fascinating why because relates knew educated explains concept research signs reasons interested important relate wondered notice perspective researching realize learn concepts science mind question reasoning educate explained caused grasp reason implications
Problem Solving and Decision Making	issues concern decisions issue unfortunately decide considering relevant difficult problem concerning future better hard focus attention plan serious choose consider choices plans concerned regarding problems decided relevance
Help & Guidance	care guide helps need helping
Hazard & Damage	damage harmful defects scary fear lose risk worse dangerous harming awful worry fight
Health	disease illnesses healthy breath health treatments treating drug vaccines unhealthy illness sweating diseases healthier doctor medical treat nutrition diet doctors
Food	foods ate eat table corn milk drink eating lunch dinner grocery turkey nutrition fruits food delicious cook hungry broccoli diet hunger
Finance	money pay buy price expensive
Farming	cows agriculture seeds crops dairy

hazards; topics such as these are likely to generalize to utility-rich writing in other subjects. The next two groups (Health and Food) are likely to be relatively specific to biology, as knowledge in this area relates to the human body and its needs for cure and nutrition. Finance and farming are also commonly mentioned in UV-rich writing on biology topics.

Features for utility value prediction

In this section, we describe the features that will be used for predicting the human-assigned utility value score of a writing sample. We developed a set of features that address the form and the content of personalized writing.

Pronouns

The qualitative analysis shows that grammatical categories that signal self, addressee, or other human reference are typical of UV-rich writing. For the following grammatical categories, we calculate log frequency per 1,000 words:

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

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)

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 (words in the G_T genre_topic set). 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. The estimation of typical genre-topic vocabulary is done on TRAIN and DEV data.

- GENREVOC: Log of the type proportion of genre_topic words out of all words in the essay.

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 (adapted from 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*. Total number of words and phrases: 144.
- 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). Total number of words and phrases: 623.

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. A heavy reliance on past test verbs is a hallmark of narrativity (or, as Biber and Conrad 2009 call it, a "reconstructed account of events", p. 240). Use of common action, mental, and desire verbs could signal sequences of actions and personal stance towards those (Biber and Conrad 2009, 68), both relevant to UV writing. We therefore define the following features (using log frequency per 1,000 words):

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

Likely UV content

Building on the qualitative observations of common UV content in the training data and on previous work by Harackiewicz et al. (2016), we capture specific content and attitude using dictionaries from LIWC (Pennebaker et al. 2015). 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. Total number of words: 1,393.
- 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*, *Female*, and *Male*. Total number of words: 756.
- Insight: Words that signify cognitive engagement, such as *think*, *know*, *consider*. Total number of words: 259.
- Health: Words that refer to matters of health and disease, such as *clinic*, *flu*, *pill*. Total number of words: 294.
- Risk: Dangers and things to avoid, such as *danger*, *doubt*. Total number of words: 103.
- Ingestion: Example words: *eat*, *dish*, *pizza*. Total number of words: 184.

Preliminary Study: Evaluating Features and Feature Combinations

For the preliminary study, we evaluated the extent to which we can predict the utility value as assigned by human raters to essays written on one of the six topics in the dataset. We calculate the correlation with UV score for each feature on its own, as well as for feature families. To build a prediction model for a feature family, we train a random forest regressor using the relevant set of features on the TRAIN data with Pearson's correlation (r) as the objective function. Table 5 shows evaluations of various features and feature families on the DEV data.

Single Features

First, we consider the correlations for the individual features. The single strongest predictor of UV score for each genre is one of the set of pronouns – first person

Table 5 Pearson correlations with UV score for various features and feature families, on DEV data

Feature Family (FEATURE)	Essay	Letter	Society	Summary
Pronouns	**.759**	**.442**	**.527**	**.544**
PRO_SG1	(s) .714	(s) .255	(s) .754	(s) .375
PRO_PL1	–.088	n.s.	n.s.	(s) .421
PRO_2	n.s.	(s) .358	n.s.	.161
DET_POS	(s) .237	(s) .299	n.s.	.153
PRO_INDEF	.078	n.s.	n.s.	.152
General Voc	**.302**	**.200**	. n.s.	**.260**
WORDLN	–.165	–.126	n.s.	–.137
WF_MEDIAN	(s) .213	.119	(s) .286	.175
ACADEMICWL	(s) –.210	–.136	n.s.	–.095
CONCRETE	.138	.168	n.s.	.135
TYPES	.187	n.s.	(s) .221	.195
GenreVoc	(s) **.219**	(s) **.378**	n.s.	(s) **.377**
ArgNarr	**.289**	**.286**	**.249**	**.195**
ARGDEV	.199	.137	n.s.	.176
HEDGEBOOST	.131	.187	n.s.	.135
COMVERBS	.175	.178	(s) .270	.138
PASTTENSEVERBS	.165	n.s.	(s) .448	n.s.
UV content	**.306**	**.313**	n.s.	**.318**
AFFECT	.109	.181	n.s.	.143
SOCIAL PROCESSES	n.s.	(s) .228	n.s.	(s) .201
INSIGHT	(s) .255	n.s.	n.s.	.081
HEALTH	.181	.111	n.s.	(s) .294
RISK	n.s.	.114	n.s.	.072
INGESTION	n.s.	.117	n.s.	n.s.
ALL	**.784**	**.543**	**.527**	**.622**
ALL WITHOUT GenreVoc	**.787**	**.500**	**.527**	**.586**
ALL WITH |*r*| >.200 (s)	**.762**	**.501**	**.584**	**.600**

For single features, Pearson correlations between feature values and UV scores are shown. For feature families (including GenreVoc, which is a single-feature family), we show correlation attained by a random forest regressor trained on the TRAIN data using the relevant set of features. Correlations that are not significant ($p > 0.05$) are marked as "n.s." For all genres apart from Society, correlations at or above 0.07 are statistically significant; for Society, it is above 0.22, due to the much smaller size of the Society dataset. The last row shows performance of the strong feature set for the given genre. The strong features are marked with (s) in the table singular for Essay and Society, second person pronouns for Letter, and first person plural pronouns for Summary. These findings are mostly consistent with expectations. In the Essay genre, students are required to express utility to themselves; in the Letters – to an addressee who is likely to be mentioned through second-person pronouns *you*, *your*, *yours*, *yourself*. In the Summary genre, which is a control genre, there are very few essays that express individual personal utility; if there is an expression of utility, it is mostly of the generic type that applies to humans at large and can be referenced as *we*, as the writer would include himself or herself in a general human reference. The finding for first person plural pronouns for the Society genre is somewhat surprising – given the instructions for the genre, one would expect that the we-as-humans reference would be more predictive of UV score, yet it is not a significant predictor. This can be explained by observing the Utility Value score distribution in the Society data shown in Table 3 – 74% of the training essays are at the utility level score of 1 (we-as-humans level), so identifying the language for this level is not very useful for attaining good correlations, it is the more individualized references corresponding to utility value score of 3 and 4 that can capture observed differences in UV scores. Another finding that is surprising at first sight – the negative correlation between the use of plural pronouns and UV score for Essays – can be likewise interpreted considering the distribution of UV scores in the Essay data. The bulk of the essays are in the top UV score categories, so extensive use of we-the-humans references that are typical of UV level 1 is actually predicting low UV, relatively for writing in this genre. As a family, PRONOUNS is the strongest predictor of UV scores across all genres.

The General Vocabulary features perform as expected, and largely consistently across the genres: writing with higher UV tends to use shorter words on average that are also less academic, more concrete, and more frequent in English at large. It is also the case that writing with high UV tends to have a larger variety of different words (TYPES).

The GenreVoc feature that captures language that is typical of the given genre and topic based on frequency analysis produces weak-to-medium correlation with score.[4] The correlations are stronger for Letter and Summary genres, since the UV score distributions there are more skewed, so typical writing in the genre coincides with high (Letter) or low (Summary) UV scores. For Essay genre, there is more variability in the UV scores, so a feature that captures typical Essay words for the given topic is less strongly correlated with UV scores.

Features capturing argumentative and narrative elements (ArgNarr family) have consistently positive weak-to-medium correlations with UV scores across genes. The narrativity features (COMVERBS and PASTTENSEVERBS) have significant correlations with UV score for Society essays, while the argument features do not; this should be interpreted with care, however, since the evaluation dataset for Society is small, with only 84 essays.

[4] For Summary, the raw correlations are negative, since language that is typical of Summary writing tends to have utility value of 0, this being the control genre. The random forest regression model has given it a negative weight, to produce a positive correlation with score.

Features that capture specific content that is expected to support UV expression (UV Content) show consistent though weak correlations with UV scores across genres. Of the six categories, Affect and Health have significant correlation with UV score across Essay, Letter, and Summary genres; Risk and Ingestion show the lowest performance, with the latter having a significant correlation with UV only in the Letter genre. None of the UV Content categories shows a significant correlation with UV score for Society, and neither does the family model that combines the six categories together.

We observe that even though the four genres are likely to have different common patterns of language use, features that correlate with UV-rich writing are relatively stable across genres. Indeed, the directions of the significant correlations are stable with a single exception of plural pronouns in Essays vs Summaries; all features apart from Ingestion are significant predictors in at least two different genres; even the magnitudes of the correlations are often quite stable (consider family-level correlations for all but Pronouns family). Thus, even though Essays, Letters, and Summaries have different mean values for Affect (2.6 for Summaries, 3.1 for Essays, 3.7 for Letters), there are correlations of 0.11-0.18 between Affect and utility values scores across the three genres. We believe these findings lend support to a hypothesis that expressions of utility value have their own characteristic linguistic patterns that interact with, but are not completely overshadowed by, effects of genre.

Combined Models

We note that for all genres apart from Society, the different features in a family combine effectively to produce correlations that are stronger than for any single feature in the family. A model that combines all features together (the “ALL” row in Table 5) outperforms all single features and single families for all genres apart from Society. Moreover, building a model that combines only the strongest features for a given genre – all features that attain an absolute correlation of at least 0.2 with the UV score – produces a model that is 2–4 points inferior to the ALL model (see “ALL WITH $|r| > 0.200$” row in Table 5). These findings suggests that features with various levels of predictive power can be combined effectively by the random forest regressor, and there is sufficient training data to learn appropriate weights for the features.

The situation is quite different with the Society set; the pattern of results shows clear signs of over-fitting the training data. Models that combine multiple features – Pronouns and ALL – perform worse by far than the single best feature PRO_SG1. Reducing the number of features and only using the stronger ones (5 features instead of 21) improves over the ALL model (0.584 vs 0.527), differently from the finding for the other genres. It seems that the small dataset size precludes the machine learned model from making an effective use of the weak features. Based on the results on the development set, we will use ALL models for Essay, Letter, and Summary genres, and PRO_SG1 feature for Society in the blind test evaluation in Experiment 1. We will also evaluate the performance of the single best feature family, PRONOUNS, on all genres.

In addition, we build a model that includes all features apart from GenreVoc, in view of an evaluation on a blind test set with new students. For this feature, the

Table 6 Confusion Matrix for the Essay model with ALL features, on DEV set (human – columns; machine – rows)

	0	1	2	3	4
0	6	3	0	0	1
1	12	72	1	2	0
2	10	16	22	18	5
3	3	12	55	266	190
4	0	1	2	43	100

TRAIN+DEV data was used in two ways: (a) to find how this feature can best contribute to the prediction of UV scores, with random forest regressor – this is similar to all features; (b) to find the genre_topic words.[5] As such, there is a possibility that the random forest regressor would over-estimate the predictive ability of this feature when it comes to score essays that were not used for either (a) or (b). We therefore build models without this feature; see the row “ALL WITHOUT GenreVoc” in Table 5. The system’s performance on Essay data is not hurt by the removal of the GenreVoc feature, while the performance on Letter and Summary data has dropped by 4 and 3.5 points, respectively. This result suggests that if the context of use is such that a large pool of essays need to be scored collectively, and information about topics and genres in the pool can be extracted ahead of scoring, then using GenreVoc feature is likely to help performance. In Experiment 1, we will evaluate the generalizations of the models with and without the GenreVoc feature.

In order to gain a better understanding of the performance of resulting models, Table 6 shows the confusion matrix for the Essay genre, using rounded predictions from the regression. The matrix suggests that the system is fairly precise in detecting essays at the lowest range – 0 and 1. Out of the total of 135 essays put by the human raters in these categories, only 16 are given a high UV score of 3 or more. Similarly, out of the 168 essays that were rated 0-2 by the system, only 26 have scores 3 or 4 assigned by the human raters. Thus, the system’s judgement of low UV is relatively trustworthy; more so than its ability to tell apart 3 from 4, or to recognize essays assigned the intermediate score of 2 by the human raters – the bulk of these essays are over-rated by the system towards the score of 3.

Experiment 1: Predicting UV Scores in Writing on Known Topics

Table 7 shows the results for the models selected using the preliminary study for the four genres, using the TEST set that contains essays by unseen students responding to one of the six topics represented in the training data in one of the four genres. We train the models on TRAIN+DEV data and evaluate on TEST. For the setting containing all

[5] We departed from the common practice of splitting the data into model training set and feature development sets and using only the feature development set for building the feature (namely, for finding genre-topic sets), since frequency-based estimations of the sets require substantial amounts of data, as does model training. We opted for utilizing all available data for feature training, model training, and the preliminary evaluation, and reserved the blind TEST set for the final evaluation.

Table 7 Experiment 1: evaluation on the blind TEST set using Pearson's r (rows 1-4) and other measures (rows 5–6)

	Essay	Letter	Society	Summary
ALL WITHOUT GENREVOC	.779	.368	.811	.607
PRO_SG1			.836	
PRONOUNS	.766	.333	.859	.574
ALL	.786	.358	.799	.633
ALL (QWK)	.717	.368	.595	.543
ALL (Spearman)	.696	.409	.818	.590

the features, we additionally report performance using quadratically weighted kappa and Spearman correlation as objective functions. These two measures complement Pearson's correlation results by focusing more on ranking (Spearman) and focusing more on chance-corrected agreement on an integer scale (quadratically weighted kappa using rounded predictions), whereas Pearson's coefficient shows the extent to which the predicted and the true values are linearly related.

For Essays, the performance of the ALL model is at or above 0.7 for all metrics, suggesting that the system is reasonably accurate in both the ranking and the alignment with the 0-4 scale. For Society, the results across the measures show much variability, with good ranking of essays but substantial mis-alignment with the 0–4 scale, reflected in the lower value of quadratically weighted kappa.

The performance of the ALL model is consistent with the preliminary evaluations for Essay and Summary genres. For these two genres, this is the best model, better than the version without GenreVoc feature, especially for Summary. This result suggests that GenreVoc features generalize quite well to unseen essays and summaries written on known genres and topics. The results for Society are generally high, with the best performance attained by PRONOUNS feature set, followed by the single PRO_SG1 feature. This affirms the earlier observation that the small training set for this genre is insufficient to build robust models with many weak features, so models with just the PRONOUNS feature set, or even just the PRO_SG1 feature, perform better than ALL.

The Letter genre is quite challenging for the NLP system, the large number of training instances notwithstanding. One of the reasons for this finding could be the skewness of the utility value scores for this genre – more than 90% of essays scored 3 or 4. In order to perform well on these genres, the automated models need to learn to distinguish between adjacent scores on the UV scale (3 vs 4), which could be a more difficult task than in the Essay context, where we observe a larger variability in UV scores. It is also possible that, differently from personal essays, where frequent reference to oneself using a first person singular pronoun is a hallmark of high UV writing, the beneficiaries of the utility value statement in Letters are more diverse – both in terms of actual referents (*you*, *family*, *children*) and in linguistic form (*children*, *offspring*, *your child* point to the same type of significant individuals but use different words).

The results for Letters are substantially worse on TEST data than those observed in the preliminary evaluation for the Letter genre (compare results in Table 7 to those in Table 5); the exclusion of GenreVoc feature does not help much. We observe that the PRONOUNS model also sustains a large drop in performance between preliminary and TEST evaluations. To better understand the reasons for this finding, we calculated the UV score distribution in TEST vs TRAIN data, and found that the distribution in TEST is even more highly skewed towards the high end of the scale than the TRAIN data: In TRAIN data, 59% received 4 and 33% received the score of 3, while in TEST data the figures are 70% and 24%, respectively. The shift in the distribution might be responsible of the discrepant performance.

While the model cannot reliably differentiate between score levels 3 and 4 for letters, it is worth finding out whether the model can reliably identify letters with no or very little utility value (0 or 1), since, in an envisioned application, a system could be used to flag writers who struggle with articulating utility. We therefore checked whether the 7 letters with utility values of 0 or 1 in the Letter TEST data were ranked at the low end of the UV scale by the system. We found it to be the case for some but not all of them: 4 out of the 7 cases fall in the lowest 10 % of the UV scores, while the other 3 instances got high UV scores from the system. The following letter received a human UV score of 0, while the model gave it the score of 3.5:

> Dear Cousin,
> I received your letter about your attempts at finding "the one." It saddens me to hear that you are having difficulty determining the type of relationship that you should be looking for. In search of a healthy relationship, I hope you can keep a certain biological relationship in mind that should point you in the right direction, and it is a relationship so small that you need a microscope to see it.
> *.... a paragraph with a technical comparison between Mitochondria and Chloroplasts ....*
> As you can see, mitochondria and chloroplasts are perfect for one another; sharing their products and resources to benefit the other is an important function in a healthy relationship. They also have some vital similarities in structure, reproduction, and history. All of the qualities found in this relationship should also be present in the relationship you reach in your quest to find "the one." I hope that you find this letter helpful and informative, and good luck with your search. **Letter, Plant Physiology, human UV = 0; system UV = 3.5**

In the letter quoted above, the connection between the real-life concern of the addressee (finding a match for a romantic relationship) and the relationship of mutual complementarity in the biological function of mitochondria and chloroplasts is that of a rather far-fetched analogy that does not seem to suggest any concrete solutions for the real-life issue. The metaphorical use of "healthy" (to describe the perceived quality of an inter-personal relationship, rather than the physical well-being of a given individual), the use of help and guidance language (helpful, point in the right direction), use of first and second person pronouns, anthropomorphic use of affective (perfect for one another) and social processes language (relationship, sharing) to discuss cells, all combine to yield a high UV score for this letter from the automated system, while the human score is 0.

On the other hand, the following letter was given a low UV score by the system (1.5), while the human-assigned label is 4, the highest utility value. The letter addresses a concern of a family member who is a hunter observing a decline in the rate of growth of the local deer population; the author explains why the observation is not necessarily a cause for concern. First, it is likely that there are not many references to hunting in the training data, hence the system might have not had a chance to have seen this as a UV expression; secondly, the letter, while mentioning family members (dad, grandpa, uncle) does not actually address the main recipient (dad) as "you", since the advice is given (indirectly) to a different individual (grandpa). We conjecture that this indirect pattern (letter to X about Y's problem) was not observed sufficiently frequently in the training data for the system to address it effectively.

> Dear Dad,
> Grandpa called me the other day and mentioned that the deer population is still growing around H, Wisconsin. However, the overall growth rate has decreased in recent years. Local deer hunters have become concerned with the decreased growth rate being related to the predation by wolves or the presence of disease such as chronic wasting disease. That made me start to think about what might be going on with the deer population. The truth of the matter is that the deer population is actually density dependent. In other words, the deer population is undergoing logistic growth.
> *... a paragraph describing logistic growth model ....*
> Therefore, hunters should be less concerned by the decreasing population growth being attributed to disease or predation. The true cause is limited resources. If hunters were to increase their success rate during the legal seasons the shift in population density will allow the deer growth rate to increase. So I would tell Uncle J to go out to the Rod & Gun and practice on a few more targets this year. **Letter, Ecology, human UV = 4; system UV = 1.5**.

Experiment 2: Predicting UV in Essays on New Topics

In order to further investigate the generalization of the automated UV scoring models, we observe that if the model is to be used for scoring expression of utility value in essays in biology in a variety of colleges and even perhaps high schools, it is likely that the specific set of topics in biology that would be addressed in the course would differ from institution to institution. We therefore ask how well the model generalizes to new topics that were not addressed in the system's training data. To simulate the new-topic context, we split the training data in a given genre by topic, train on 5 of the topics, and test on the 6th. The data sizes for the evaluations are shown in Table 8.

We evaluate two kinds of models for the cross-topic evaluation. One is a no-semantics model that uses only the Pronoun feature set (**Pron**). We expect this model to show robust performance across topics, as these features use functional, rather than content words; it is the latter that are most likely to change with topic, while the former should not depend on topic (much). The second model is a model with all features apart from GenreTopic (**Full**). The reason for the exclusion of the GenreTopic

Table 8 Sizes of datasets, per genre per topic

Topic	Essay	Letter	Society	Summary
Plant Physiology	492	673	70	757
Animal Physiology	648	512	29	256
Genetics	660	584	61	816
Cell biology	443	402	66	582
Ecology	602	432	32	685
Evolution	400	389	27	597

These are the sizes of the test data for the given topic and genre. For example, a model that would attempt to predict UV scores in Essays on Evolution will be trained on 2,845 samples in the Essay genre responding to all topics other than Evolution, and tested on 400 Essay samples on Evolution

feature is that it cannot be calculated for samples written in response to a new topic – recall that it relies on genre_topic sets that are estimated based on frequency of word usage in samples responding to the same topic across different genres. All models are built using random forest regression, as before. Table 9 shows the results.

First, we observe that for the genre with the most promising known-topic performance, Essay, the generalization to new topics is good, with performance meeting or exceeding $r = 0.7$ for every single new topic. This is an encouraging finding. On the other hand, the Letter genre that was the most difficult for the system in

Table 9 Experiment 2: Pearson's r between the model's UV score and the human-assigned UV score, for new-topic evaluation

Topic	Essay		Letter		Society		Summary	
	Pron	Full	Pron	Full	Pron	Full	Pron	Full
Plant Physiology	.752	.766	.435	.421	.835	.835	.542	.548
Animal Physiology	.735	.755	.286	.404	.444	.444	.657	.674
Genetics	.714	.741	.341	.368	.429	.429	.510	.539
Cell biology	.728	.737	.499	.554	.522	.421	.546	.571
Ecology	.745	.747	.417	.457	.698	.647	.512	.529
Evolution	.687	.701	.386	.398	.609	.609	.544	.528
Average New-Topic	.727	.741	.394	.434	.590	.564	.552	.565
Average Known-Topic (from Table 5)	.759	.787	.442	.500	.527	.527	.544	.586
Std.	.024	.022	.075	.066	.157	.165	.075	.066
Max minus Min	.065	.065	.213	.186	.406	.414	.147	.146

Full model is significantly better than Pron model, using Wilcoxon Signed Ranked test over 24 pairs of values (6 topics X 4 genres). Test statistics: W = 110, n = 20, p (2-tail) < 0.05. The last two rows show measures of variation across the scores for the different topics in the given genre – standard deviation (Std) and maximum score minus minimum score (Max minus Min). Row "Average Known-Topic (from Table 5)" shows performance on preliminary known-topic evaluations; the results are copied from Table 5 for ease of reference the known-topic setting has sustained the largest relative and absolute drop in performance between known-topic and new-topic settings (.500 vs .434). Clearly, the current feature set does not handle Letters satisfactorily.

Next, we observe that the Full model consistently outperforms the Pron model on all sets apart from Society; it is also not the case that the performance of the Pron model is more consistent across topics than that of Full model (see the last two rows in Table 9). For Society, the models either have the same performance (because the Pron feature set is driving the performance) or the Full model performs worse (Cell Biology and Ecology). It seems that for cases where these is little training data available (between 215 and 258 for the different topics in Society set), the model with only the pronoun features and, generally, with fewer features, would be expected to perform in a more robust fashion. If a substantial amount of data is available for training (the order of magnitude of 2,500 essays), the Full model is clearly preferable, both in known-topic and in a new-topic setting, but one might want to opt for a simpler model if training data is scarce.

Related Work

Harackiewicz et al. (2016) performed an exploratory analysis of writing in UVI genres versus control (Summary) writing, using a subset of categories from Linguistic Inquiry and Word Count (LIWC) dictionary (Pennebaker et al. 2015). Hypothesizing that UVI writing would be more personal and social and hence contain references to self and other people, they selected the following LIWC categories: personal pronouns (I, us, your), long words (> 6 letters), family words (mom, sister), friend words (friend, neighbor), human words (adult, baby), social processes words (advice, discuss, encourage). Further, they hypothesized that the process of making connections between course material and own life is expected to deepen the student's cognitive involvement with the material, that would be reflected in increased use of words from the cognitive mechanism dictionary (cause, conclude, explain), insight words (consider, idea, understand), and causal words (because, effect, hence). The authors found that UVI writing had significantly higher proportions of 8 out of the 9 categories; no significant effect was observed for causal words. The results of our qualitative analysis second the observations in Harackiewicz et al. (2016), as both identified the use of people-centered language as being characteristic of UV writing, as well as words that express improved understanding.

Stark et al. (2012) addressed the question of classifying phone interactions into types, such as business, residential, family, unfamiliar. They hypothesized that people who share close social ties would tend to engage in conversations on a wider variety of topics than business conversations or conversations with strangers. To measure language variety, they computed language entropy over unigram word distribution of each conversation in the dataset; the results supported the hypothesis. Stark et al. (2012) also used LIWC categories to aid classification; categories of pronouns, insight, and cognitive mechanism were among the categories that distinguished conversations with family members from conversations with other people as well as conversations with family members from residential conversations with non-family. The alignment of these findings with ours and those of Harackiewicz et al. (2016) suggests that themes of cognitive engagement, improved understanding, and reference to self and others mark personal or even intimate communication.

Pérez-Rosas and Mihalcea (2014) addressed the question of differentiating deceptive from truthful short essays on the topics of abortion, death penalty, and feeling towards best friends collected from writers belonging to three cultures (US, India, Mexico). The authors hypothesized that psychological categories like those captured in LIWC would help discriminate between deceptive and truthful statements. Analyzing the specific categories that received high weights in the models, they observe that the categories of reference to self, first person plural, and friends are strong predictors of truthful language, while second person references and references to humans at large were predictive of deceptive language. Furthermore, the category of insight came out as the most predictive category of the truthful class in US data. These results indicate that when people are asked to write truthfully, they opt for highly personalized writing with language suggesting cognitive engagement. Authentic, truthful, and personal engagement being the goal of utility value interventions, perhaps a UVI genre that allows reference to individuals beyond the intimate circle of self, friends, and family should be treated with some caution, since people might find it easier to write without true and authentic engagement about strangers.

NLP methods have been successfully applied to assessing writing quality related to writing proficiency, such as analysis of discourse structure (Falakmasir et al. 2014; Persing et al. 2010; Burstein et al. 2003), rating of discourse coherence quality (Burstein et al. 2013a; Somasundaran et al. 2014; Yannakoudakis and Briscoe 2012; Miltsakaki and Kukich 2004), detection of errors in English conventions (Leacock et al. 2014), analysis of appropriateness of vocabulary choice (Beigman Klebanov and Flor 2013) and of use of sources (Rahimi et al. 2014; Beigman Klebanov et al. 2014). Automated writing evaluation systems are widely used to score writing item responses on high-stakes assessments (Burstein et al. 2013b; Foltz et al. 2013), to evaluate students' written work in writing instruction applications (Burstein et al. 2004), and to evaluate the quality of reviewer comments in student peer review systems (Xiong et al. 2012).

Above and beyond evaluating the written quality of a response (e.g., adherence to English conventions, discourse coherence), 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; Pérez-Rosas and Mihalcea 2014; Hu and Liu 2004; Ranganath et al. 2009; Neviarouskaya et al. 2010; Beigman Klebanov et al. 2010; Beigman Klebanov et al. 2008; Greene and Resnik 2009; Pedersen 2015; Resnik et al. 2013; Mulholland and Quinn 2013).

Utility value intervention can be thought of as a writing-to-learn activity, that affords students the opportunity of a deeper engagement with the content of the course while considering its personal relevance. This activity is related to what Grossman (2008) termed *content-based reflective writing*; there, students of psychology were encouraged to tutor in an elementary school and reflect, in detail, on the relationship between concepts from a course and their tutoring experience, in order to attain a deeper understanding of the course material. The focus in Grossman's intervention was on a detailed point-to-point comparison between theoretical notions and definitions and their manifestations in the tutoring practice; in UVI, the students are not requested to undergo a particular kind of practical experience (such as tutoring), but to connect the content of the course to a remembered experience (such as experience of illness, or a conversation with a friend who is facing a health-related or professional problem) or an imagined experience (such as what working in a particular profession would entail and how to prepare for it). Beauchamp and Thomas (2010) and Conway (2001) consider envisioning future professional identity an integral part of reflective practice.

Discussion

The results of our study are, we believe, encouraging for the potential application of UV scoring technology to help scale up the UV intervention, especially in the personal Essay genre. In particular, we found that the automated system can score UV with high correlations with human-provided scores, even in a context where the topic of the essay has not been seen in the training data. Moreover, the confusion matrix analysis suggests that most confusions are between the top two UV scores (3 and 4), while identification of essays with low UV can be done with good accuracy. In one application, an automatically generated UV score could be used to identify students who failed to include UV in their essays in order to direct them to a conversation with the instructor about possible applications of the acquired knowledge. Another useful result of the current study is the semi-automatic analysis of the kinds of themes that come up in UV-rich writing; a further development of UV-analysis technology could enable retrieval of UV excerpts using theme index to a database of UV-rich writing, as a preparatory activity for the student when thinking about his own utility.

The UV score distributions (see Table 3) suggest that the Letter genre is more conducive to UV-rich writing, as fewer than 10% of Letters have a UV score below 3 on the 0-4 scale, compared to 28% of personal Essays. Nevertheless, it is not clear that one would necessarily recommend that UVI be cast in the Letter genre – both on the grounds of variability in assignments (in cases where the intervention is administered multiple times, such as after each new module), and taking into account the findings of the deception studies such as Pérez-Rosas and Mihalcea (2014) that found that people tend to use second-person terms when writing non-truthfully. From the point of view of computational modeling for prediction of UV scores, the Letter genre is more challenging, due both to the skewness of UV distribution and to a more diverse patterns of writing-to-address-other-people's-problems, as compared to reflecting on one's own plans and goals. In particular, we observed cases of overly general, philosophical advice to others (think about the mutually beneficial relation between mitochondria and chloroplasm when looking for a significant other in your life), as well as cases where the addressee of the letter was not actually the person whose problem is being discussed. The Letters genre clearly stands to benefit from further NLP work of finding additional robust indicators of UV-rich writing, as well as markers of contemplative, rather than solution-focused, writing.

For the personal Essay genre, the score distribution is more spread out, and the correlations attained by the computational model are generally high, $r = .78 - .79$. Moreover, the model can generalize to new, completely unseen topics in biology, while still retaining the $r > .7$ correlations with score. We found, however, that fairly strong performance can be obtained by a very simple feature that log transforms the number of observed first personal singular pronouns in the essay. This finding suggests that technology-delivered and technology-evaluated utility value intervention could be prone to "gaming" by students by inserting a large number of "I"'s into their essay. To counteract such eventuality, we believe it would be possible to estimate a reasonable range of the extent of reference to self in an essay, and have the system flag excessive self-reference (students might know that using "I" is good, but, when doing this mechanically rather than in the context of actual reflective writing, it might be hard to guess how much is too much). Other means of counteracting simple-minded gaming could be letting the students know that the teacher will be scoring a certain (small) proportion of the essays, chosen at random; this might deter students form following simplistic gaming strategies. Development of additional indicators of UV-rich writing would likewise help make the UV assessment system more robust to construct-irrelevant behavior on the part of the student.

The literature on writing-to-learn suggests that asking students to write with a particular audience in mind can alter the nature and quality of the resulting writing in a way that relates to learning. In science learning context specifically, Gunel et al. (2009) found that high school students who wrote an explanation of a biology concept with a peer or a younger student in mind performed significantly better on conceptual questions than students writing for the teacher or the parents. Among the current UVI versions, Letter contains an explicit element of audience design, in that students are asked to address the letter to a family member or close friend, while the target audience was not clearly specified in the other versions of UVI. It is possible that the more "lay" nature of the addressee generally prompted writing that was less academic, more concrete, and more affective (as in "Dear Auntie, I hope your knees do not hurt anymore"); since UVI writing is characterized by some of the same elements across the genres (see Table 5), UV writing in the Letter genre might have been more congruent and easier than in Essay and Society tasks. Evaluating the effect of target audience on UV articulation is an interesting area for future work.

Summary and Conclusion

Studies in social psychology have shown that consideration of ways in which the STEM material relates to the student's personal and social life and values can enhance interest and performance in the STEM subject, as well as lead to improved motivation reflected in better retention in STEM majors. To our knowledge, this is the first application of NLP technology for predicting utility value expressed in a student's writing sample. The results are encouraging, especially for the genre of personal essay, while the letter genre is more challenging for the NLP system reported in this paper.

The ability to accurately assess the expressed utility value in a writing sample opens up the possibility of scaling the UVI through automated administration and scoring. Automated scoring is particularly important since (a) biology teachers often view this kind of assignment as tangential to their main goal of teaching the subject matter of the STEM course, and (b) human scoring of writing samples for expression of utility value requires significant training. Analyses such as the qualitative investigation reported here could also be utilized to design scaffolds for utility-value writing, namely, guided exercises aimed at helping students articulate the utility value they derive from studying the STEM subject. Finally, the cross-genre analysis of utility-value writing conducted here sheds light on linguistic characteristics of this type of writing that recur across different genres (personal essay, letter, summary) as well as those that are specific to certain genres; these, along with a thematic breakdown of commonly used utility language, could aid teachers when designing and evaluating utility-value writing assignments in biology, as well as serve as a comparison when investigating utility value articulation in other disciplines.

Our future work includes development of more sophisticated NLP models for predicting utility values, as well as evaluation of the generalizations of the models to writing in other STEM disciplines.

References

Abouelenien, M., Perez-Rosas, V., Mihalcea, R., & Burzo, M. (2014). Deception detection using a multimodal approach. In *Proceedings of the 16th ACM international conference on multimodal interaction* (pp. 58–65). New York: ACM.

Aull, L.L., & Lancaster, Z. (2014). Linguistic markers of stance in early and advanced academic writing: a corpus-based comparison. *Written Communication*, *31*, 151–183.

Beauchamp, C., & Thomas, L. (2010). Reflecting on an ideal: student teachers envision a future identity. *Reflective Practice*, *11*, 631–643.

Beigman Klebanov, B., Beigman, E., & Diermeier, D. (2010). Vocabulary choice as an indicator of perspective. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 253–257). Uppsala, Sweden: Association for Computational Linguistics.

Beigman Klebanov, B., Diermeier, D., & Beigman, E. (2008). Automatic annotation of semantic fields for political science research. *Journal of Information Technology and Politics*, *5*(1), 95–120.

Beigman Klebanov, B., & Flor, M. (2013). Word association profiles and their use for automated scoring of essays. In *Proceedings of the 51st annual meeting of the association for computational linguistics* (pp. 1148–1158). Sofia, Bulgaria: Association for Computational Linguistics.

Beigman Klebanov, B., Madnani, N., Burstein, J., & Somasundaran, S. (2014). Content importance models for scoring writing from sources. In *Proceedings of the 52nd annual meeting of the association for computational linguistics* (pp. 247–252). Baltimore, MD: Association for Computational Linguistics.

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge, UK: Cambridge University Press.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Brewer, C., & Smith, D. (2011). Vision and change in undergraduate biology education: a call to action. http://visionandchange.org/files/2011/03/Revised-Vision-and-Change-Final-Report.pdf.

Brown, E., Smith, J., Thoman, D., Allen, J., & Muragishi, G. (2015). From bench to bedside: a communal utility value intervention to enhance students' biomedical science motivation. *Journal of Educational Psychology*, *107*(4), 1116–1135.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: the criterion service. *AI Magazine*, *25*(3), 27–36.

Burstein, J., Kukich, K., Wolff, S., Lu, J., & Chodorow, M. (1998). Enriching automated essay scoring using discourse marking. In *Proceedings of the ACL workshop on discourse relations and discourse marking* (pp. 15–21). Montréal, Canada: Association for Computational Linguistics.

Burstein, J., Marcu, D., & Knight, K. (2003). Finding the write stuff: automatic identification of discourse structure in student essays. *IEEE Intelligent Systems*, *18*(1), 32–39.

Burstein, J., Tetreault, J., & Chodorow, M. (2013a). Holistic discourse coherence annotation for noisy essay writing. *Dialogue and Discourse*, *4*(2), 34–52.

Burstein, J., Tetreault, J., & Madnani, N. (2013b). The e-rater® automated essay scoring system. In Shermis, M., & Burstein, J. (Eds.) *Handbook of automated essay scoring: current applications and future directions*. New York: Routledge.

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

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

Conway, P.F. (2001). Anticipatory reflection while learning to teach: from a temporally truncated to a temporally distributed model of reflection in teacher education. *Teaching and Teacher Education*, *17*, 89–106.

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

Durik, A.M., & Harackiewicz, J.M. (2007). Different strokes for different folks: how personal interest moderates the effects of situational factors on task interest. *Journal of Educational Psychology*, *99*, 597–610.

Eccles, J. (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.

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

Falakmasir, M.H., Ashley, K.D., Schunn, C.D., & Litman, D.J. (2014). Identifying thesis and conclusion statements in student essays to scaffold peer review. In *Proceedings of the 12th international conference on intelligent tutoring systems* (pp. 254–259). Honolulu, Hawaii: Springer International Publishing.

Foltz, P., Streeter, L., Lochbaum, K., & Landauer, T. (2013). Implementation and application of the intelligent essay assessor. In Shermis, M., & Burstein, J. (Eds.) *Handbook of automated essay evaluation: current applications and new directions* (pp. 68–88). New York: Routhledge.

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

Greene, S., & Resnik, P. (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* (pp. 503–511). Boulder, Colorado: Association for Computational Linguistics.

Grossman, R. (2008). Structures for facilitating student reflection. *College Teaching*, *57*, 15–22.

Gunel, M., Hand, B., & McDermott, M.A. (2009). Writing for different audiences: effects on high-school students' conceptual understanding of biology. *Learning and Instruction*, *19*(4), 354–367.

Gunel, M., Hand, B., & Prain, V. (2007). Writing for learning in science: a secondary analysis of six studies. *International Journal of Science and Mathematics Education*, *5*, 615–637.

Harackiewicz, J., Canning, E., Tibbetts, Y., Priniski, S., & Hyde, J. (2016). Closing achievement gaps with a utility-value intervention: Disentangling race and social class. *Journal of Personality and Social Psychology*, *111*(5), 745–765.

Harackiewicz, J., Durik, A., Barron, K., Linnenbrink-Garcia, L., & Tauer, J. (2008). The role of achievement goals in the development of interest: reciprocal relations between achievement goals, interest, and performance. *Journal of Educational Psychology*, *100*, 105–122.

Harackiewicz, J., Tibbetts, Y., Canning, E., & Hyde, J. (2014). Harnessing values to promote motivation in education. In Karabenick, S., & Urden, T. (Eds.) *Advances in motivation and achievement* (pp. 71–105). Bingley, UK: Emerald Group Publishing Limited.

Hidi, S., & Harackiewicz, J.M. (2000). Motivating the academically unmotivated: a critical issue for the 21st century. *Review of Educational Research*, *70*, 151–179.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 168–177). Seattle, Washington: ACM.

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

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

Hulleman, C.S., Durik, A.M., Schweigert, S.A., & Harackiewicz, J.M. (2008). Task values, achievement goals, and interest: an integrative analysis. *Journal of Educational Psychology*, *100*, 398–416.

Leacock, C., Tetreault, J., Gamon, M., & Chodorow, M. (2014). *Automated grammatical error detection for language learners*, 2nd edn. Morgan & Claypool Publishers: San Rafael, CA.

Mihalcea, R., & Strapparava, C. (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* (pp. 309–312). Singapore: Association for Computational Linguistics.

Miltsakaki, E., & Kukich, K. (2004). Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering*, *10*, 25–55.

Mulholland, M., & Quinn, J. (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* (pp. 680–684). Nagoya, Japan: Asian Federation of Natural Language Processing.

NCES (2013). NCES 2013-152: STEM in postsecondary education: entrance, attrition, and coursetaking among 2003-04 beginning postsecondary students. http://nces.ed.gov/pubs2013/2013152.pdf.

NCES (2014). NCES 2014-001: STEM attrition: college students' paths into and out of STEM fields. http://nces.ed.gov/pubs2014/2014001rev.pdf.

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2010). Recognition of affect, judgment, and appreciation in text. In *Proceedings of the 23rd international conference on computational linguistics* (pp. 806–814). Beijing, China: COLING 2010 Organizing Committee.

PCAST (2012). Engage to excel: producing one million additional college graduates with degrees in science, technology, engineering, and mathematics. https://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-engage-to-excel-final_2-25-12.pdf.

Pedersen, T. (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* (pp. 46–53). Denver, Colorado: Association for Computational Linguistics.

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

Pennebaker, J., Boyd, R., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin.

Pérez-Rosas, V., & Mihalcea, R. (2014). Cross-cultural deception detection. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 440–445). Baltimore, Maryland: Association for Computational Linguistics.

Persing, I., Davis, A., & Ng, V. (2010). Modeling organization in student essays. In *Proceedings of the 2010 conference on empirical methods in natural language processing, EMNLP '10* (pp. 229–239). Stroudsburg, PA, USA: Association for Computational Linguistics.

Pintrich, P. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology*, *95*(4), 667–686.

Prain, V., & Hand, B. (2016). Coming to know more through and from writing. *Educational Researcher*, *45*, 430–434.

Rahimi, Z., Litman, D.J., Correnti, R., Matsumura, L.C., Wang, E., & Kisa, Z. (2014). Automatic scoring of an analytical response-to-text assessment. In *12th international conference on intelligent tutoring systems (ITS)* (pp. 601–610). Honolulu, Hawaii: Springer International Publishing.

Ranganath, R., Jurafsky, D., & McFarland, D. (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* (pp. 334–342). Singapore: Association for Computational Linguistics.

Resnik, P., Garron, A., & Resnik, R. (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* (pp. 1348–1353). Seattle, Washington, USA: Association for Computational Linguistics.

Ripley, B.D. (1996). *Pattern recognition and neural networks*. Cambridge: Cambridge University Press.

Smith, J., Brown, E., Thoman, D., & Deemer, E. (2015). Losing its expected communal value: how stereotype threat undermines women's identity as research scientists. *Social Psychology of Education*, *18*, 443–466.

Somasundaran, S., Burstein, J., & Chodorow, M. (2014). Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of the 25th international conference on computational linguistics* (pp. 950–961). Dublin, Ireland: The COLING Organizing Committee.

Stark, A., Shafran, I., & Kaye, J. (2012). Hello, who is calling?: can words reveal the social nature of conversations? In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 112–119). Montréal, Canada: Association for Computational Linguistics.

Wigfield, A. (1994). Expectancy-value theory of achievement motivation: a developmental perspective. *Educational Psychology Review*, *6*, 49–78.

Xiong, W., Litman, D., & Schunn, C. (2012). Natural language processing techniques for researching and improving peer feedback. *Journal of Writing Research*, *4*(2), 155–176.

Yannakoudakis, H., & Briscoe, T. (2012). Modeling coherence in ESOL learner texts. In *Proceedings of the 7th workshop on building educational applications using NLP* (pp. 33–43). Stroudsburg, PA, USA: Association for Computational Linguistics.