“Hello, [REDACTED]”: Protecting Student Privacy in Analyses of Online Discussion Forums

Nigel Bosch
University of Illinois at Urbana–Champaign
Champaign, IL, USA
pnb@illinois.edu

R. Wes Crues
University of Illinois at Urbana–Champaign
Champaign, IL, USA
rwcrues@gmail.com

Najmuddin Shaik
University of Illinois at Urbana–Champaign
Champaign, IL, USA
shaik@illinois.edu

Luc Paquette
University of Illinois at Urbana–Champaign
Champaign, IL, USA
lpaq@illinois.edu

ABSTRACT

Online courses often include discussion forums, which provide a rich source of data to better understand and improve students’ learning experiences. However, forum messages frequently contain private information that prevents researchers from analyzing these data. We present a method for discovering and redacting private information including names, nicknames, employers, hometowns, and contact information. The method utilizes set operations to restrict the list of words that might be private information, which are then confirmed as private or not private via manual annotation or machine learning. To test the method, two raters manually annotated a corpus of words from an online course’s discussion forum. We then trained an ensemble machine learning model to automate the annotation task, achieving 95.4% recall and .979 AUC (area under the receiver operating characteristic curve) on a held-out dataset obtained from the same course offered 2 years later, and 97.0% recall and .956 AUC on a held-out dataset from a different online course. This work was motivated by research questions about students’ interactions with online courses that proved unanswerable without access to anonymized forum data, which we discuss. Finally, we queried two online course instructors about their perspectives on this work, and provide their perspectives on additional potential applications.

Keywords

Text anonymization; discussion forums; online learning.

1. INTRODUCTION

Online education is an essential part of many university programs [12] and has many potential benefits, such as convenience, scalability, and lower cost for both students and institutions. However, personal connections and discussions with fellow students could be quite negatively impacted if there are no opportunities for students to interact with each other as they can easily do in face-to-face classes. Hence, many online courses include optional or required discussion forums, in which students can talk about course content or connect with each other. For researchers, the textual contents of these forums is a valuable source of knowledge for understanding more deeply how students experience learning in online environments (see studies such as [4, 6, 8, 11, 16, 18, 23, 26, 37]). A significant barrier to analyzing the contents of these forums is the private nature of information students can and do disclose to each other, such as names, affiliations, locations, and contact information. Analyzing these data often requires anonymization before researchers can ethically and legally access the data for analyses. In this paper, we propose and evaluate a method specifically designed for anonymizing student-generated text in discussion forums.

There are various types of identifying information students share on discussion forums. Some are relatively straightforward to remove, such as phone numbers and email addresses, which follow a relatively limited set of formatting patterns. Others are less predictable – especially the names of people and places, which can appear in various forms (e.g., nicknames), overlap with dictionary words (e.g., May, Lane, Bob), or refer to entities not listed in course rosters (e.g., family members, locations). For example, one student in data we analyzed posted potentially identifying information about a pet:

> “Hello [REDACTED], I am also a pet lover. I have a [REDACTED] schnauzer, whose name is [REDACTED]. What's your work at the dog kennel? How many puppies are there in the kennel? It seems lots of fun!”

While other students refer to themselves or others by alternate names, as in the case of this student:

> “Hi guys,My name is [REDACTED], but I prefer to be called [REDACTED]. I was born and grew up in [REDACTED], but I moved to [REDACTED] when I was in 7th grade.”

Moreover, students frequently misspell identifying and non-identifying information (e.g., “*Battlestar Gallactica*”, “*When we icnrease entropy does it change delta G as well?*”), which – combined with grammatical errors – resulted in relatively poor anonymization quality in our early efforts built on named entity

Nigel Bosch, Wes Crues, Najmuddin Shaik and Luc Paquette ""Hello, [REDACTED]": Protecting Student Privacy in Analyses of Online Discussion Forums" In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 39 - 49

recognition software. Hence, we sought more robust methods to detect identifying information to be redacted.

Identifying information can occur in forums when students organize study groups, address questions and answers to each other, and other situations. Students may receive meaningful benefits from exposing private information online – for example, if it enables them to connect more closely with peers they may never meet offline. Examples such as those above are especially common in introductory discussion forums at the beginning of courses, where students get to know each other. However, the presence of identifying information prevents researchers at many academic institutions offering online courses from analyzing forum data (at least without individual permission from each student), and thus from enhancing student learning experiences through applications of research. We focus on this problem for the specific case of university-level online courses, of which there are many, and propose an automated text anonymization solution that rivals human accuracy, despite the variance in form, content, and spelling inherent in student-generated text.

1.1 Privacy Concepts and Anonymization Strategies for Text

There is a large body of previous research on removing identifying information from text. A primary focus of prior work has been specifically on removing names and identifying information from medical records (see [24] for a review). One of the earliest methods employed a template-matching approach to find names, addresses, phone numbers, and other identifying information in medical records (e.g., notes written by doctors) [35]. Later research with similar methods has shown that template-matching approaches can be quite accurate in held-out (unseen) medical records data, achieving a recall of .943 [28], which compared favorably to inter-human agreement on the same data.

Early work on anonymizing text also led to the concept of *k*-anonymity [33, 34], in which a formal guarantee is made for a particular dataset that every person in that dataset is indistinguishable from *k*-1 or more other people in the dataset. This has resulted in additional text anonymization research that goes beyond names of people and places to include identifying characteristics such as specific diseases and treatments that may be sufficiently unique to reduce *k* with some effort [3]. In general, these works utilize lists of known names and forms of names (e.g., "Dr. [name]") to identify words for removal in text – forms which are used infrequently in online course discussion forums – and tend to focus on the unique needs of medical literature anonymization.

Named entity recognition (NER) is another closely-related field that focuses on finding *and classifying* names in text [25]. Modern NER approaches typically rely on machine learning to discover names in text by learning from large corpora of annotated or partially-annotated text. In theory, NER can be applied for anonymization purposes by finding names and removing or replacing those from classified categories of interest (i.e., people, places, and organizations that may be employers) [10]. However, modern NER systems are typically trained on large amounts of data that differs considerably from discussion forum data (e.g., the entire contents of Wikipedia), and do not generalize well to new domains [20, 21].

Previous research has also studied privacy and anonymity in structured data (e.g., directed graphs, tabular records) that is relevant to forum anonymization. For example, social network analysis shows that individuals in one social network can be identified in another network based on who they interact with [27], which might occur across course discussion forums. The network of semantic and stylistic relationships between words can also identify individuals from text data [2, 5]. Such connections have led to the concept of differential privacy. Differential privacy is one of the strongest types of data privacy [14], which guarantees that it is impossible to determine whether or not a query individual's data was included in a given dataset or result. While we do not propose providing such a strong guarantee for anonymizing discussion forum text for research analyses – given the need for obfuscating much of the text that could be needed for analyses (e.g., person-specific sentiment words) – we instead propose a set of goals that allow well-intentioned researchers to access data with minimal exposure to identifying information.

1.2 Novelty of the Problem

Our method for automatically anonymizing discussion forum text aims to satisfy several goals needed for practical application. Specifically, the automatic method should:

1) Achieve accuracy similar to human accuracy, if it is to be used as a replacement for manual annotation
2) Not require annotation of large amounts of domain-specific text data for development or validation
3) Not rely on lists of student names, which may be unavailable (as was the case in our work), may not capture the diversity in naming conventions of students from various cultures, and may not capture nicknames frequently used by students

Approaches relying on NER methods satisfy goals 1 and 3 well, but not goal 2. Conversely, approaches developed for the needs of medical record anonymization typically satisfy goal 1 and potentially goal 2 (though this has not been well tested and may not be the case if the style of medical text differs notably from online discussion forum text), but typically do require information such as lists of individual's names, and do not satisfy goal 3.

The approach we propose here satisfies the three goals outlined above, utilizing a set-theoretic approach to drastically reduce the burden of manual annotation and machine learning to further automate the manual annotation process.

2. ANONYMIZATION METHOD

The anonymization procedure consists of three broad steps (see Figure 1 for an overview). First, we extract a set of possible name words from the discussion forum text. Second, we classify possible names as either actual names or not names, via manual annotation or machine learning. Third, we remove the identified names from text, along with other likely identifying information that can be found via regular expressions, including emails, URLs, and phone numbers.

2.1 Data Collection

We obtained discussion forum text data from two online courses offered at a large, public university in the Midwestern United States. The first course (*course 1*) was an elective STEM course offered using the Moodle learning management system [13], while the second course (*course 2*) was an introductory STEM course that was required for students in some majors and was offered using the LON-CAPA learning management system [22]. Discussion forum participation was a required, graded component of both courses, and students were quite active in the forums. We obtained two semesters of course 1 data separated by two years and one semester

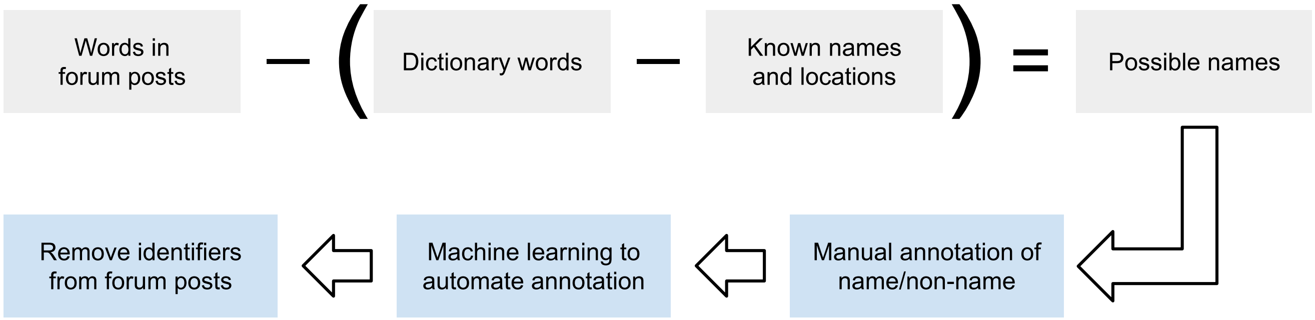
Words in forum posts
−
(
Dictionary words
−
Known names and locations
)
=
Possible names
Manual annotation of name/non-name
Machine learning to automate annotation
Remove identifiers from forum posts

Figure 1. Anonymization method overview. Grey boxes indicate data, blue boxes indicate processing steps. Minus signs indicate set subtraction.

of data for course 2. In the first semester of course 1 there were 14,082 posts made by 226 individuals – including the instructor, whose identity and forum posts also need to be anonymized. In the second semester of course 1 there were 9,217 posts made by 295 individuals, and in course 2 there were 930 posts made by 78 individuals. Forum activities consisted of personal introductions, questions about and discussions of topics in the course, team formation/coordination for group projects, and others.

We developed the anonymization method by examining the largest dataset (the first semester of course 1, referred to as *training data*) and utilized the remaining two datasets as completely unseen held-out testing data. The second semester of course 1 (referred to as *holdout 1*) provides a test of generalization of the method over time, while course 2 (referred to as *holdout 2*) serves as a test with a different course topic, learning management system, and instructor.

We obtained approval from our institutional review board and the instructors of the courses before collecting and analyzing data. However, we were only permitted to access anonymized data for analyses. Hence, we developed the anonymization method in cooperation with university data warehouse staff, who ran code for analyses on original forum data and shared the anonymized results.

2.2 Narrowing the List of Possible Names

There are several possible categories for each word (sequence of consecutive non-whitespace characters, after removing punctuation) in a discussion forum post. The word may be an identifying name referring to a person or a place, an English word[1], a misspelling, or a non-English word (e.g., numbers, other languages). The most challenging and time-consuming aspect of anonymization is to determine whether a particular word is identifying or not. We applied a set-theoretic approach to drastically reduce the scope of the problem, narrowing down the list of all words in forum posts to a small subset of possible names, which are then much less time-consuming to annotate.

The top row of Figure 1 illustrates the possible name extraction process. We started with a dictionary of over 100,000 English words [36], including common loanwords, and removed any words that overlapped with a list of over 23,000 cities, political regions, and countries (words such as South, New, etc. that were part of place names)[2]. We then also removed any words from the dictionary that overlapped with a list of over 7,000 first and last names obtained from U.S. census data. Thus, the dictionary contained only words that were not the names of people or places – words like *wormhole* and *dalliance*, but not *so* or *will*. We then removed these non-name dictionary words from the list of all unique words in discussion forum posts, leaving only possible names.

2.3 Feature Extraction

We extracted various features to help both human annotators and machine learning models classify each possible name as a name or non-name word. Features can be categorized into two basic types: densely-distributed *ad hoc* features and sparsely-distributed word presence features. Ad hoc features calculated for each possible name consisted of:

- Count of occurrences
- Word index in the first post where the word was used
- Count of words in the first post where the word was used
- Proportion of occurrences where the word was capitalized
- Proportion of occurrences where the word was at the beginning of a sentence
- Proportion of mid-sentence occurrences (not at the beginning of a sentence) where the word was capitalized
- Proportion of occurrences where the word was mid-sentence and capitalized
- Whether the word was a dictionary word or not (before modifying the dictionary)
- Whether the word was in the U.S. census list of first/middle names
- Whether the word was in the U.S. census list of last (family) names
- Frequency of the word in the U.S. census list of first/middle names

[1] This paper focuses on English-language text. However, the method could be repeated for other languages by replacing English-specific components (e.g., the dictionary) with another language.

[2] Obtained from http://www.geonames.org

- Frequency of the word in the U.S. census list of last names
- Whether the word was in the list of world cities
- Whether the word was in the list of political regions (e.g., states, territories)
- Whether the word was in the list of countries
- Count of dictionary words that were within one edit (deletion, insertion, replacement, or transposition) of the word
- Count of dictionary words that were within two edits of the word

The list of ad hoc features resulted from several rounds of error analysis and iterative refinement, which was necessary to reach classification accuracies comparable to human raters. Feature development proceeded approximately in order of complexity. Our original features consisted of the simplest ideas such as the count of occurrences. Complex features, such as the proportion of capitalized mid-sentence occurrences, resulted from examining prediction errors from models with simpler features. While this process may have resulted in over-fitting the feature extraction process to training data, we made no adjustments to features for final evaluation on the holdout datasets (see description of dataset annotation below).

Word presence features indicated the presence or absence of a particular word within the 10 most common words preceding the possible name word, which we refer to as context words, among all of its occurrences across forum posts. We tracked the most common context words separately for capitalized mid-sentence occurrences, capitalized occurrences, and all occurrences. This separation helps to determine whether common dictionary words like "hope" were also names. Word presence features consisted of a 1 if a particular word appeared in the ten most common context words for the possible name in question, and a 0 otherwise. Word presence features captured things like if a word was preceded by "hi" or "hello" – words which tended to indicate the presence of a name. We limited these features to the 25 most common overall context words, yielding 75 total context word presence features (since there were 3 capitalization conditions). Additionally, we included an "other" count category for all less-common context words, yielding another three features (one for each capitalization condition). For example, the words "tea" and "coffee" might occur among the 10 context words for a particular possible name, but be too infrequent across all possible names to rank in the top 25; we would thus count these both as "other" and calculate features for the number of "other" words in the context words for that particular name, the number of capitalized "other" words, and the number of "other" words capitalized in the middle of a sentence. Thus there were 95 features in total: 17 densely-distributed and 78 sparsely-distributed.

2.4 Manual Annotation of Possible Names

Two raters iteratively annotated possible names derived from the training data, checked agreement, and updated an annotation scheme to resolve patterns of common disagreement. Annotators had access to features listed above, as well as the possible names themselves. They did not, however, have access to the actual forum posts nor to associated possible name pairs (first and last names together), thereby mitigating unnecessary exposure to possible identifying information.

In the first round, raters annotated 200 randomly-selected possible names as either names, non-names, or unknown. Of these 200, they annotated 10 as unknown. The annotation guide was subsequently revised to remove the unknown category, since ultimately a name/non-name decision must be made for anonymization, and to clarify unknown cases. Unknowns primarily consisted of famous individuals' names (e.g., Obama), which we classified as names out of an abundance of caution. For the remaining 190 cases, the raters achieved 87.4% agreement and Cohen's κ = .734 (confidence interval = [.634, .833]).

Raters annotated a different set of 200 randomly-selected possible names in the second round to test the updated annotation guide. They achieved 89.5% agreement and κ = .773 (confidence interval = [.681, .865]). After this round we added the mid-sentence capitalization features described above, to help disambiguate disagreements noted by the raters.

Raters completed a third round of annotation to test the final annotation guide, achieving 92.7% agreement and κ = .842 (confidence interval = [.820, .864]) on all 2,588 instances in the training data, indicating excellent agreement [7]. After this round they also annotated a sample of 650 randomly-selected possible names from the holdout 1 dataset, though we removed 50 of these when we later discovered that they were erroneously included due to UTF-8 encoding issues. This left a holdout sample of 600 possible names, which we deemed sufficient to produce a tight confidence interval for agreement, given the confidence intervals previously obtained with just 200 possible names. On the holdout 1 dataset the raters achieved 93.8% agreement, with κ = .864 (confidence interval = [.823, .907]), indicating that they were able to apply the annotation guide to a new dataset with at least as much agreement as the original dataset. Finally, raters discussed each of their disagreements to reach a definitive name/non-name label for each of the 600 possible name instances in holdout 1.

A single rater annotated a sample of 600 possible names in the holdout 2 dataset as well. Given the excellent agreement between raters, we deemed a single rater sufficient for this task. Specifically, the more conservative rater (higher recall; see rater comparison in Table 1 results) annotated the holdout 2 dataset.

The final data thus consisted of 2,588 labeled instances in the training dataset (35.5% annotated as names), 600 in holdout 1 (36.0% annotated as names), and 600 in holdout 2 (44.5% annotated as names), which we used to train and validate the automatic name classification procedure.

2.5 Name/Non-name Classification

The process of extracting possible names greatly reduces the burden of manual annotation and limits raters' exposure to identifying information. We sought to further reduce these concerns by automating the classification step.

We evaluated two quite different machine learning approaches and ultimately combined them via decision-level fusion. The first classification algorithm was *Extra-Trees* [15], which is a variant of Random Forest that trains multiple trees (500 in our case) based on random subsets of data, and adds further randomness by choosing random points at which to divide the data in feature space. Extra-Trees makes no strict assumptions about data distribution, and thus works well for the features in this paper, which include densely- and sparsely-distributed features with vastly different ranges and distributions. Moreover, Extra-Trees has inherent feature selection (dimensionality reduction) capabilities, since irrelevant features can simply be ignored when constructing each tree. We utilized the implementation of Extra-Trees available in *scikit-learn* for this model [30].

The second approach we evaluated was a deep neural network (DNN) implemented with *TensorFlow* using the stochastic gradient descent optimizer [1]. We developed a custom structure for the DNN to suit the specific properties of the problem (Figure 2). The feature space is relatively large (95 dimensions) for a model with no inherent dimensionality reduction capabilities, so we added regularization to constrain model complexity. The densely- and sparsely-distributed features call for different regularization methods, however. Several of the densely-distributed features were highly correlated, and the number of features (17) was relatively small. Thus, we applied L2 regularization for densely-distributed features [29]. Conversely, we applied L1 regularization to the sparsely-distributed features, of which there were many (78), since L1 pushes the weight of irrelevant features toward 0. We then concatenated the post-regularization outputs of fully-connected layers for densely- and sparsely-distributed features, and stacked additional fully-connected layers (which were regularized via dropout [31]). Finally, we added a fully-connected sigmoid activation output layer (i.e., logistic regression) to predict name or non-name.

We evaluated models via nested four-fold cross-validation on the training dataset. In this approach, we randomly selected 75% of instances (possible names), trained a model on those instances, and tested it on the remaining 25% of instances. We repeated the process three more times so that each instance was in the testing set exactly once. During training, we weighted false negative errors (incorrectly classifying a name as a non-name) twice as heavily as false positive errors (incorrectly classifying a non-name as a name), since we were more concerned about missing identifying information than about accidentally removing non-identifying words. False positive errors might adversely affect some analyses (e.g., if the word “joy” was mistaken for a name, thereby changing the result of sentiment analysis), but would not harm student privacy.

We tuned hyperparameters (model settings) for both models via nested cross-validation, in which we tested different hyperparameters and selected the best combination of hyperparameters based on cross-validated mean squared error. Note that this step took place nested within training data only, via 4-fold cross-validation within the training data of the outer 4-fold cross-validation loop, so that hyperparameters were not selected based on test set accuracy.

For the Extra-Trees model, we tested hyperparameters consisting of the minimum number of instances required for each leaf of the tree (values of 1, 2, 4, 8, 16, or 32) and the maximum proportion of features to consider when creating each tree branch (values of .25, .5, .75, or 1.0). For the DNN, we searched hyperparameters including the number of neurons in each hidden layer (2, 4, 8, 16, or 32), L2 regularization strength (.1, .01, or .001), L1 regularization strength (.1, .01, or .001), dropout regularization strength (0, .25, or .5), number of hidden layers after the concatenation layer (0, 1, 2, or 4), and the learning rate (.01, .001, or .0001). The hyperparameter search space consisted of the cross product of these values (i.e., grid search). Hence, training was time-consuming (several days), but the trained models can be applied to an entire course’s data in less than 10 seconds.

Finally, we re-trained the models on all training data and applied to the held-out dataset. We then combined model predictions to form a decision-level fusion model by simply averaging Extra-Trees and DNN predictions, for both cross-validated training set predictions and holdout predictions.

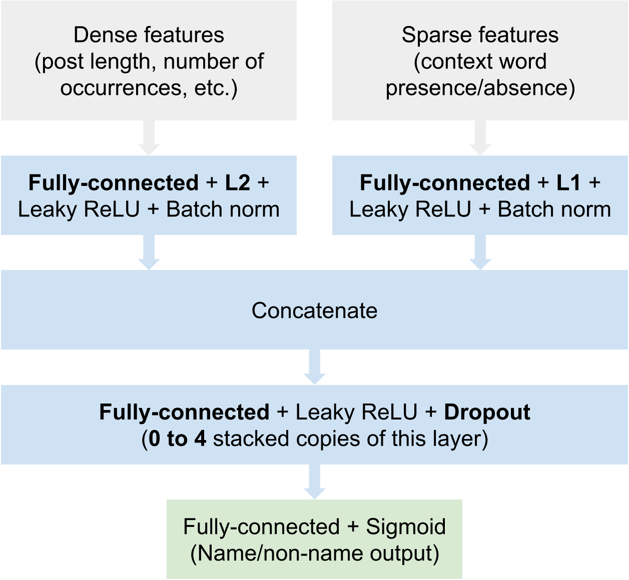
Dense features
(post length, number of occurrences, etc.)
Sparse features
(context word presence/absence)
Fully-connected + **L2** + Leaky ReLU + Batch norm
Fully-connected + **L1** + Leaky ReLU + Batch norm
Concatenate
Fully-connected + Leaky ReLU + **Dropout**
(**0 to 4** stacked copies of this layer)
Fully-connected + Sigmoid
(Name/non-name output)

Figure 2. Custom DNN structure. Elements marked in bold were hyperparameters (number of neurons, regularization strength, or number of layers) tuned via nested cross-validation.

2.6 Removal of Identifiers

After names have been identified via manual annotation or automatic classification, removal of identifying information is relatively straightforward. First, we removed potential identifying information that follows known regular expression patterns, consisting of email addresses, URLs, phone numbers, and other numbers – which we removed in case they might represent things like social security numbers or other identifiers. As noted in previous research [35], such information can be identified with essentially 100% accuracy via pattern matching, and the challenging cases are names, nicknames, misspellings, and other name variants that we focus on in this paper. We replaced each pattern match with a placeholder (e.g., *phone_placeholder*) so that they could potentially serve as context words for name classification or be measured during analyses of forum content. Second, we replaced all identified name words with a placeholder, regardless of capitalization (see Error Analysis sections below for measures of how much this may result in non-name words being accidentally removed).

3. CLASSIFICATION RESULTS

We evaluated machine learning results in terms of common accuracy metrics below, but also compared human raters to evaluate the utility of the automatic approach as a replacement for manual annotation.

3.1 Machine Learning Accuracy

Table 2 contains key results of the automatic name classification method. Overall, results show that the models were highly accurate, reaching area under the receiver operating characteristic curve (AUC) as high as .981 in cross-validated evaluation on the training data, .979 on the holdout 1 dataset, and 956 on the holdout 2 dataset. AUC ranges from 0 to 1, where 1 indicates perfect classification

Table 2. Name vs. non-name classification results. FN indicates false negatives (names classified as non-names) and FP indicates false positives (non-names classified as names). Precision and recall refer to the positive (name) class. Acc refers to the percentage correctly classified.

Model	FN	FP	AUC	Acc	Cohen's κ	Precision	Recall	Base rate	N
Cross-validated training data									
Extra-Trees	52	202	.971	90.2%	.793	.811	.943	.355	2588
Custom NN	37	157	.981	92.5%	.841	.849	.960	.355	2588
Fusion	37	173	.980	91.9%	.828	.836	.960	.355	2588
Holdout course 1 (later semester)									
Extra-Trees	10	56	.976	89.0%	.772	.786	.954	.360	600
Custom DNN	11	55	.975	89.0%	.771	.788	.949	.360	600
Fusion	10	54	.979	89.3%	.778	.792	.954	.360	600
Holdout course 2 (different course)									
Extra-Trees	11	54	.950	89.2%	.784	.826	.959	.445	600
Custom DNN	16	61	.950	87.2%	.744	.805	.940	.445	600
Fusion	8	54	.956	89.7%	.794	.827	.970	.445	600

accuracy, 0 indicates completely incorrect classification, and .5 indicates random chance level. Thus, these results indicate models were accurate and generalized well from training data to the holdout data from 2 years later as well as the holdout data from another course.

We evaluated models with several different classification metrics in an effort to uncover any particular ways in which the method might be failing, but our primary concern was the number of false negatives – that is, the number of names misclassified as non-names. The decision-level fusion model yielded the lowest number of false negatives in each dataset (37, 10, and 8 for training, holdout 1, and holdout 2 respectively), and thus we intend to apply this model for practical use, though both of the individual models exhibited high accuracy as well.

The machine learning results also compare favorably to examples of previous work on anonymization of medical literature. In one study [28], researchers reported .967 recall on a training dataset (versus .960 for our decision-level fusion model) and .941 recall in a holdout testing set (versus recalls of .954 and .970 for our fusion model across the two holdout datasets). Moreover, our method did not require a dataset-specific list of names, as is common in previous work. While results are not exactly comparable, since base rates and predicted rates may have differed, they are strongly indicative of similar accuracy.

3.2 Comparison to Human Raters

Measuring annotation agreement between two human raters is one way to determine how "difficult" a task is, and whether a machine learning solution is close in accuracy. We computed machine learning model accuracy by comparing predictions to the resolved set of labels produced by raters; here we compare raters (pre-resolution) to each other. Since neither rater necessarily represents the ground truth more than the other, we computed comparison metrics alternately treating each rater as the ground truth. Table 1 shows these results computed with the same accuracy metrics as the machine learning model, using the holdout 1 dataset.

Results show that the machine learning method was close to, and in some respects equally as accurate as the human raters. Recall and false negatives (FN) are especially important to consider for minimizing the risk of identifying information being revealed, and both showed that the fusion machine learning model (recall = .954, FN = 10) was close to or better than human accuracy depending on

Table 1. Details of human raters' agreement, treating each rater as ground truth individually to allow comparison to machine learning accuracy on the same task. AUC refers to the minimum proper AUC (calculated via linear interpolation with a single point) because raters provided only yes/no annotations, not probabilities.

	Ground truth	
	Rater 1	**Rater 2**
FN	30	7
FP	7	30
AUC	.945	.923
Acc (% agreed)	93.8%	93.8%
Cohen's κ	.864	.864
Precision	.864	.964
Recall	.964	.864
Base rate	.360	.360
N	600	600

which rater we considered as ground truth (recall .864 or .964, FN = 30 or 7). Note, though, that for cases where algorithmic recall exceeds human rater recall, human rater precision is correspondingly better since it is reciprocal with recall when treating a single rater as ground truth. In terms of κ, human raters do seem likely to be superior to the machine learning fusion model (κ = .864 versus .778). Thus, for some sensitive applications of the method, human raters may be needed. However, the fusion model primarily makes false positive (FP) errors, which are less of a privacy concern than FN errors.

The difference in FN between raters (FN = 7 versus 30) also indicates that there was some inconsistency in terms of tendency of one rater versus the other to make a classification of name or non-name. This is one potential advantage of making continuous-valued predictions (probabilities, in this machine learning case) of name versus non-name, because it allows setting a threshold. For human raters, thresholds are implicit but not easily or specifically controllable. As noted previously, we weighted false negative errors twice as heavily as false positive errors, though that is a parameter that could be adjusted for the particular needs of a dataset.

4. HOLDOUT 1 (LATER SEMESTER) ERROR ANALYSIS

We conducted an analysis of cases where machine learning predictions were incorrect, focusing on the decision-level fusion model applied to the holdout 1 dataset. Analysis of false negatives is important to discover the severity of cases where names are left unredacted, while analysis of false positives is important to quantify the amount of text that will be unnecessarily anonymized (replaced with placeholders).

4.1 False Negatives

False negatives are the most serious errors, since they may result in identifying information being revealed. There were 10 false negatives, which we examined to determine how serious these errors might be and to determine why they might have occurred. Human raters disagreed on 6 of the 10 words (and they only disagreed 37 times total – see Table 1), and only agreed to classify those 6 as names after discussing. This indicates that these were exceptionally difficult cases, even for humans. Furthermore, the machine learning method made similar false negative errors as human raters.

Of the 10 false negatives, there were 2 dictionary words (“long” and “mercy”), which may have indeed not been names. One of the 10 was the name of an entertainment company, which may have been an identifying characteristic (an employer) or, more likely, simply a reference to entertainment. Similarly, one was a name from a famous television show, and one was the name of a U.S. national park. The remaining words included a concatenated combination of words that was likely a filename but could have been a username, two non-English words, one name that seems likely to be a person’s first name (though it appeared only once in a forum post and was not capitalized), and one possible last name.

In sum, while there were several false negative predictions, examination of these cases reveals that even human raters initially disagreed for most of them and that it is quite possible that most of them, except probably the apparent last name, are indeed not names.

4.2 False Positives

While false positive errors are less serious, since they do not compromise identity, they do pose a challenge to subsequent analysis of forum text if important words are removed (e.g., words that might indicate sentiment, like “joy”).

The decision-level fusion model made 54 false positive errors. We observed several broad categories that capture most of these instances. First, we observed several geographical regions (e.g., “Africa”, “European”) that were too broad for our definition of identifying information – which was restricted to political regions – or even extraterrestrial (e.g., “Ganyemede”). Second, there were misspellings (e.g., “hellium” instead of “helium”), most of which were correctly identified as non-names but a few of which were not. Third, there were abbreviations such as “NBA” and “DOI”. Fourth, there were references to popular culture, such as “Overwatch” and “Kerbal”, which are indeed names but not identifying information. Finally, there were several domain-specific words, which we do not include as examples to avoid unintentional identification of the course from which data were collected.

Among these false positives, the most commonly-occurring word occurred just 26 times in 9,217 posts (the total size of the dataset from which holdout 1 data were sampled), most occurred only once, and all false positives combined appeared 191 times in those posts. This indicates that even though some non-name words were mistakenly removed from posts, the impact on the overall text was minimal.

5. HOLDOUT 2 (NEW COURSE) ERROR ANALYSIS

We performed similar analyses of classification errors for the holdout 2 dataset. However, it was not possible to compute inter-rater disagreement for the misclassified cases in holdout 2 because only one rater performed annotations.

5.1 False Negatives

There were just 8 FN errors among the 600 possible names in the holdout 2 dataset. Of these eight, three were abbreviations for university-specific terms, including a building name, a college (collection of university departments) name, and the name of a major. A further three FN errors were slang terms for large metropolitan areas with populations over 4 million. One was half of a misspelled two-word city name, and the last was a local street name.

None of the FN errors in this dataset were student names. The most serious errors are perhaps the university-specific terms, which could narrow down the identity of students when combined with other factors. However, in isolation (or even combined with each other) these terms match hundreds or thousands of students, and thus do not pose a likely risk for researchers hoping to analyze forum data.

5.2 False Positives

There were 54 FP errors in the holdout 2 dataset, which differ somewhat from the FP errors observed in holdout 1. Course 2, from which holdout 2 data were collected, utilized Roman numerals for assignment numbers, which were frequently mistaken for names. Additionally, the domain-specific content of course 2 required students to discuss a large number of letter combinations (strings) that do not represent words, and which were also often mistaken for names.

Like holdout 1, there were several misspellings mistaken for names in holdout 2 results. For example, "callender" (calendar), "hewlp" (help), and "ssolid" (solid) were FP errors. However, these account for very few redactions since these misspellings occurred only infrequently. The most notable FP was the word "my", which occurred 293 times in the 930 posts in the holdout 2 dataset. Surprisingly, "My" was capitalized in 32.1% of its occurrences, including 15.4% of occurrences in the middle of sentences. This was somewhat unexpected, but appears to have frequently occurred when students did not punctuate the end of sentences and instead used line breaks (which we did not consider as end-of-sentence markers) to separate sentences.

Human intervention after the automatic classification step can easily correct false positive errors such as "my", however. To facilitate this, our anonymization software produces a list of names identified by the machine learning fusion model as an intermediate step before the names are removed. The list includes the fusion model's probability as well as the number of occurrences of each word in the original discussion forum data, sorted in descending order by occurrences. Researchers (or authorized staff in charge of anonymizing data) can thus easily examine the top of the list and delete any rows that are clearly high-impact FP errors before proceeding to the last step where names are redacted.

6. APPLICATIONS FOR INSTRUCTORS

Our primary motivation for developing this method was to enable research on the text of computer-mediated discussions students have with each other during their online learning experiences. However, such research may support the needs of course instructors as well, either directly via analysis methods they can easily apply, or via generalizable insights that can be applied to their courses. We thus sought out instructors of online university courses (who were not involved in the forum anonymization work or the courses analyzed in this paper) to gain better insight into instructor perspectives on scalable analysis of online course discussion forums. Specifically, we asked two instructors "*as an online course instructor, can you imagine any analyses of discussion forum text that would be informative for you?*"

6.1 Instructor 1

The first instructor was a male computer science faculty member with 14 years of university-level teaching experience, who had taught for-credit online courses at the university level as well as massive open online courses (MOOCs). He noted:

> Specifically for all courses that I don't teach I don't have a legitimate need to know that student X is enrolled in course Y. Anonymization gives us a way to easily share forum discussions between different instructors of the same course, or across department etc. And there are numerous reasons why this is useful.
>
> * Potential for early detection of struggling students and the underlying cause. (Lack of time? interest? pre-reqs? effective strategies?)
>
> * Identification of hardest components of a course.
>
> * Research projects that look at common forum post across multiple courses. e.g. fresh/ sophomores/ seniors.

He also noted that when working with students to improve courses it is necessary to have anonymized data:

> If I want to give the data to an undergrad staff for analysis for course improvement purposes (rather than for research publication), I'd require that they had anonymized data.

Additionally, instructor 1 conducts and publishes research on his own courses, and offered research questions and ideas he would like to pursue that would require anonymization. These included:

> It may be possible to detect themes and generate hypotheses by skim reading the posts, but it is much harder to identify trends and quantitative trends (e.g. are there more X in the later part of the course). Also a general skimread of the forums will miss correlations with other data (e.g. students with background X tend to post more Y)
>
> Can we identify when a course pace is too fast? Compared to assuming too much prior knowledge?
>
> Suppose we consider a student's forum post action as an active intervention created by the student to affect on their own learning trajectory. How effective are these interventions? Do they also help similarly students that just read the discussion thread (and never need to post a similar issue themselves). Are they too late? Are they too early?

In sum, instructor 1 was enthusiastic about the prospect of being able to quickly anonymize online course discussion forums, and proposed several ways in which anonymization would benefit both teaching and research.

6.2 Instructor 2

The second instructor was a female statistics instructor and graduate student, with six years of university-level teaching experience. Her online courses are large, and thus provide unique challenges for teacher–student engagement. As she noted:

> This semester there are about 1,400 students enrolled in [course information redacted]. It would be beneficial for me as an instructor to have some sort of automated analysis that told me which forums and topics were getting the most activity. That would help me know which forums to look at or have my undergrad course assistants look at and answer some of the questions. It would also be beneficial because if there was a lot of confusion about a certain topic, I would know I need to re-explain that topic in lecture.
>
> I think something that identified negative words would be helpful too for the same reasons. If there's a lot of negativity on a thread- it's probably best that I go over that concept again in class to clarify any confusion.

While some of the needs noted by instructor 2 do not require access to the forum text itself (such as tools to measure forum activity), others would require researchers and developers to have access to anonymized forum text. For example, developing and validating methods for automatic assessment of confusion in forums is only possible with access to text data. Moreover, these needs highlight the difficulty of effectively utilizing online discussion forums with very large numbers of students, and the potential for automated tools to assist instructors in these courses.

7. DISCUSSION

In this study we were interested in enabling analysis of online discussion forums in university courses through removal of identifying information, even in cases where capitalization, grammar, and spelling may be unpredictable. Our results showed that automatic anonymization is possible, and that it rivals human accuracy.

In this section we discuss implications of the results for various stakeholders, including users of the anonymization method (e.g., researchers, teachers) and students whose data is subject to analyses.

7.1 Implications for Users of the Anonymization Method

The proposed anonymization method offers two main advantages to users. First, it drastically reduces workload relative to approaches like manual identification and removal of identifying information directly from the forum text. Moreover, such manual anonymization is often intractable for users because they cannot access non-anonymized data in the first place. Second, it reduces users' exposure to identifying information. Users may either utilize the machine learning approach to avoid all involvement with identifying information, or annotate possible names manually – in which case they are still protected from seeing identifiers in the full context of the original text.

Anonymization is essential in many cases for researchers to either validate existing methods or develop new methods. For example, when automatically detecting sentiment from text with tools such as *SEANCE* (Sentiment Analysis and Social Cognition Engine; [9]) it is helpful to match sentiment to forum posts to obtain examples of the context in which sentimental language occurs. It is especially important to preserve student privacy in research that requires detailed reading of forum posts. For example, domain experts might annotate and evaluate the depth of questions students ask, or the responses they receive, to answer research questions about the relationship between a student's engagement with their peers and their status as a member of demographic groups that are traditionally-underrepresented in postsecondary education.

Finally, one important consideration for applications is how well the machine learning model is likely to generalize. We showed excellent generalization across time (2 years), as well as to a new course topic, instructor, and learning management system. While the change in topic (and instructor-specific course setup) did result in different types of errors, overall accuracy remained similar. However, we did not test across university populations. Students at other universities may have different backgrounds that influence how they interact with each other or with technology, and the vernacular language they use. Moreover, the same method could be applied to anonymize student-generated text in other contexts, such as college admissions essays [32], where students may reveal identifying information but in different (non-conversational) circumstances. Thus, for generalization to a notably different context, such as a different university or type of text, we recommend annotating a testing set of possible names to validate accuracy.

7.2 Implications for Students

The objective of our method is to minimize the potential for negative impacts on student privacy introduced by analyses of unstructured student-generated text. It is important, however, to recognize that such analyses carry inherent risk even with a (hypothetical) perfectly-accurate anonymization method. For example, students might mention their involvement in a particular course in venues such as Twitter, Reddit, Facebook or others [39]. They may even post similar questions on course forums and public forums, or relate events that took place on course forums. It is unreasonable to expect perfect anonymization. Thus, it is important to take appropriate steps to limit public exposure to student data – even anonymized data – and to ensure that students reap benefits of analyses conducted on their data.

Positive impacts for students largely consist of 1) improvements made to future courses, and 2) additional capabilities afforded to instructors, both informed by research made possible through access to anonymized data. For example, researchers may be able to provide guidance to students about how to ask questions to elicit the most helpful responses. Or, as instructor 2 noted above, it might be possible to direct the attention of teaching assistants to students or topics where it is most needed.

Benefits to students are indirect in nature, and, in the case of research-informed changes to online courses, benefits might be more for future students than for the students from whom data were collected. Thus, more research is needed to sample student perspectives regarding analysis of their forum data, as well as their perspectives on the importance and impact of anonymization.

8. CONCLUSION

Access to discussion forum data is essential for researchers to better understand the experiences of students interacting with each other in web-based learning environments. However, access to these forum data is often hampered by important privacy concerns. Our approach for automatic anonymization of these data helps to resolve this issue, and has already enabled in-depth examination of forum posts [17–19, 38]. We plan to make our anonymization software publicly available[3], and hope that it will be instrumental in advancing researchers' and teachers' knowledge of student experiences, and, ultimately improving learning in online classrooms.

9. ACKNOWLEDGMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A180211 to Board of Trustees of the University of Illinois. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

10. REFERENCES

[1] Abadi, M. et al. 2016. TensorFlow: A system for large-scale machine learning. *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)* (2016), 265–283.

[2] Anandan, B. and Clifton, C. 2011. Significance of term relationships on anonymization. *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 03* (USA, Aug. 2011), 253–256.

[3] Anandan, B., Clifton, C., Jiang, W., Murugesan, M., Pastrana-Camacho, P. and Si, L. 2012. t-Plausibility: Generalizing words to desensitize text. *Transactions on Data Privacy*. 5, (2012), 505–535.

[3] See https://ilearn.illinois.edu for anonymization software

[4] Beckmann, J. and Weber, P. 2015. Cognitive presence in virtual collaborative learning: Assessing and improving critical thinking in online discussion forums. *Proceedings of the 2015 International Conference on E-Learning* (2015), 51–58.

[5] Chakaravarthy, V.T., Gupta, H., Roy, P. and Mohania, M.K. 2008. Efficient techniques for document sanitization. *Proceedings of the 17th ACM conference on Information and knowledge management* (New York, NY, Oct. 2008), 843–852.

[6] Chen, B., Chang, Y.-H., Ouyang, F. and Zhou, W. 2018. Fostering student engagement in online discussion through social learning analytics. *The Internet and Higher Education*. 37, (Apr. 2018), 21–30. DOI:https://doi.org/10.1016/j.iheduc.2017.12.002.

[7] Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. 20, (1960), 37–46. DOI:https://doi.org/10.1177/001316446002000104.

[8] Crossley, S., Paquette, L., Dascalu, M., McNamara, D.S. and Baker, R.S. 2016. Combining click-stream data with NLP tools to better understand MOOC completion. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (New York, NY, USA, 2016), 6–14.

[9] Crossley, S.A., Kyle, K. and McNamara, D.S. 2017. Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior Research Methods*. 49, 3 (Jun. 2017), 803–821. DOI:https://doi.org/10.3758/s13428-016-0743-z.

[10] Cumby, C. and Ghani, R. 2011. A machine learning based system for semi-automatically redacting documents. *Proceedings of the Twenty-Third Innovative Applications of Artificial Intelligence Conference* (Aug. 2011).

[11] Davis, G.M., Wang, C. and Yuan, C. 2019. N-gram graphs for topic extraction in educational forums. *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)* (2019), 532–535.

[12] Deming, D.J., Goldin, C., Katz, L.F. and Yuchtman, N. 2015. Can online learning bend the higher education cost curve? *American Economic Review*. 105, 5 (May 2015), 496–501. DOI:https://doi.org/10.1257/aer.p20151024.

[13] Dougiamas, M. and Taylor, P. 2003. Moodle: Using learning communities to create an open source course management system. (2003), 171–178.

[14] Dwork, C. 2008. Differential privacy: A survey of results. *Theory and Applications of Models of Computation* (Berlin, Heidelberg, 2008), 1–19.

[15] Geurts, P., Ernst, D. and Wehenkel, L. 2006. Extremely randomized trees. *Machine Learning*. 63, 1 (Apr. 2006), 3–42. DOI:https://doi.org/10.1007/s10994-006-6226-1.

[16] Harrak, F., Bouchet, F., Luengo, V. and Bachelet, R. 2019. Automatic identification of questions in MOOC forums and association with self-regulated learning. *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)* (2019), 564–567.

[17] Henricks, G.M., Perry, M. and Bhat, S. in press. Gender and gendered discourse in two online STEM courses. *Proceedings of the 14th International Conference on Learning Sciences (ICLS 2020)* (Nashville, TN, in press).

[18] Huang, E., Valdiviejas, H. and Bosch, N. 2019. I'm sure! Automatic detection of metacognition in online course discussion forums. *Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction (ACII 2019)* (Piscataway, NJ, 2019), 241–247.

[19] Jay, V., Henricks, G.M., Bosch, N., Perry, M., Bhat, S., Williams-Dobosz, D., Angrave, L. and Shaik, N. in press. Online discussion forum help-seeking behaviors of students underrepresented in STEM. *Proceedings of the 14th International Conference on Learning Sciences (ICLS 2020)* (Nashville, TN, in press).

[20] Jiang, R., Banchs, R.E. and Li, H. 2016. Evaluating and combining named entity recognition systems. *Proceedings of the Sixth Named Entity Workshop, joing with 54th ACL* (2016), 21–27.

[21] Kleinberg, B., Mozes, M., Arntz, A. and Verschuere, B. 2018. Using named entities for computer-automated verbal deception detection. *Journal of Forensic Sciences*. 63, 3 (2018), 714–723. DOI:https://doi.org/10.1111/1556-4029.13645.

[22] Kortemeyer, G., Albertelli, G., Bauer, W., Berryman, F., Bowers, J., Hall, M., Kashy, E., Kashy, D., Keefe, H., Behrouz, M.-B., Punch, W.F., Sakharuk, A. and Speier, C. 2003. The learning online network with computer-assisted personalized approach (LON-CAPA). *Computer Based Learning in Science (CBLIS 2003)* (2003), 119–130.

[23] Lee, S.Y., Chae, H.S. and Natriello, G. 2018. Identifying user engagement patterns in an online video discussion platform. *Proceedings of the 11th International Conference on Educational Data Mining (EDM 2018)* (2018), 363–368.

[24] Meystre, S.M., Friedlin, F.J., South, B.R., Shen, S. and Samore, M.H. 2010. Automatic de-identification of textual documents in the electronic health record: A review of recent research. *BMC Medical Research Methodology*. 10, 1 (Aug. 2010), 70. DOI:https://doi.org/10.1186/1471-2288-10-70.

[25] Nadeau, D. and Sekine, S. 2007. A survey of named entity recognition and classification. *Lingvisticæ Investigationes*. 30, 1 (Jan. 2007), 3–26. DOI:https://doi.org/10.1075/li.30.1.03nad.

[26] Nanda, G. and Douglas, K.A. 2019. Machine learning based decision support system for categorizing MOOC discussion forum posts. *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)* (2019), 619–622.

[27] Narayanan, A. and Shmatikov, V. 2009. De-anonymizing social networks. *2009 30th IEEE Symposium on Security and Privacy* (Piscataway, NJ, May 2009), 173–187.

[28] Neamatullah, I., Douglass, M.M., Lehman, L.H., Reisner, A., Villarroel, M., Long, W.J., Szolovits, P., Moody, G.B., Mark, R.G. and Clifford, G.D. 2008. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*. 8, 1 (Jul. 2008), 32. DOI:https://doi.org/10.1186/1472-6947-8-32.

[29] Ng, A.Y. 2004. Feature selection, L1 vs. L2 regularization, and rotational invariance. *Proceedings of the Twenty-first International Conference on Machine Learning (ICML 2004)* (New York, NY, 2004), 78–85.

[30] Pedregosa, F. et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*. 12, (Nov. 2011), 2825–2830.

[31] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*. 15, 1 (2014), 1929–1958.

[32] Stone, C., Quirk, A., Gardener, M., Hutt, S., Duckworth, A.L. and D'Mello, S.K. 2019. Language as thought: Using natural language processing to model noncognitive traits that predict college success. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (Tempe, AZ, USA, Mar. 2019), 320–329.

[33] Sweeney, L. 2002. Achieving k-anonymity privacy protection using generalization and supression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*. 10, 05 (Oct. 2002), 571–588. DOI:https://doi.org/10.1142/S021848850200165X.

[34] Sweeney, L. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*. 10, 5 (2002), 557–570. DOI:https://doi.org/10.1142/S0218488502001648.

[35] Sweeney, L. 1996. Replacing personally-identifying information in medical records, the Scrub system. *Proceedings of the AMIA Annual Fall Symposium* (Philadelphia, PA, 1996), 333–337.

[36] Ubuntu – Details of package wamerican in bionic: 2017. *https://packages.ubuntu.com/bionic/wamerican*. Accessed: 2018-06-08.

[37] Uijl, S., Filius, R. and Ten Cate, O. 2017. Student interaction in small private online courses. *Medical Science Educator*. 27, 2 (Jun. 2017), 237–242. DOI:https://doi.org/10.1007/s40670-017-0380-x.

[38] Valdiviejas, H. and Bosch, N. in press. Using association rule mining to uncover rarely occurring relationships in two university online STEM courses: A comparative analysis. *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)* (in press).

[39] Wu, J.-Y., Hsiao, Y.-C. and Nian, M.-W. 2018. Using supervised machine learning on large-scale online forums to classify course-related Facebook messages in predicting learning achievement within the personal learning environment. *Interactive Learning Environments*. 28, 1 (Sep. 2018), 1–16. DOI:https://doi.org/10.1080/10494820.2018.1515085.