Research Article

De-Identification of Laboratory Reports in STEM

Alex Rudniy, *University of Scranton*

Structured Abstract

- **Background:** Employing natural language processing and latent semantic analysis, the current work was completed as a constituent part of a larger research project for designing and launching artificial intelligence in the form of deep artificial neural networks. The models were evaluated on a proprietary corpus retrieved from a data warehouse, where it was extracted from MyReviewers, a sophisticated web application purposed for peer review in written communication, which was actively used in several higher education institutions. The corpus of laboratory reports in STEM annotated by instructors and students was used to train the models. Under the Common Rule, research ethics were ensured by protecting the privacy of subjects and maintaining the confidentiality of data, which mandated corpus de-identification.
- **Literature Review:** De-identification and pseudonymization of textual data remains an actively studied research question for several decades. Its importance is stipulated by numerous laws and regulations in the United States and internationally with HIPAA Privacy Rule and FERPA.
- **Research Question:** Text de-identification requires a significant amount of manual post-processing for eliminating faculty and student names. This work investigated automated and semi-automated methods for de-identifying student and faculty entities while preserving author names in cited sources and reference lists. It was hypothesized that a natural language processing toolkit and an artificial neural network model with named entity recognition capabilities would facilitate text processing and reduce the amount of manual labor required for post-processing after matching essays to a list of users'

names. The suggested techniques were applied with supplied pre-trained models without additional tagging and training. The goal of the study was to evaluate three approaches and find the most efficient one among those using a users' list, a named entity recognition toolkit, and an artificial neural network.

- **Research Methodology:** The current work studied de-identification of STEM laboratory reports and evaluated the performance of the three techniques: brute forth search with a user lists, named entity recognition with the OpenNLP machine learning toolkit, and NeuroNER, an artificial neural network for named entity recognition built on the TensorFlow platform. The complexity of the given task was determined by the dilemma, where names belonging to students, instructors, or teaching assistants must be removed, while the rest of the names (e.g., authors of referenced papers) must be preserved.
- **Results:** The evaluation of the three selected methods demonstrated that automating de-identification of STEM lab reports is not possible in the setting, when named entity recognition methods are employed with pre-trained models. The highest results were achieved by the users' list technique with 0.79 precision, 0.75 recall, and 0.77 *F*1 measure, which significantly outweighed OpenNLP with 0.06 precision, 0.14 recall, and 0.09 *F*1, and NeuroNER with 0.14 precision, 0.56 recall, and 0.23 *F*1.
- **Discussion:** Low performance of OpenNLP and NeuroNER toolkits was explained by the complexity of the task and unattainability of customized models due to imposed time constraints. An approach for masking possible de-identification errors is suggested.
- **Conclusion:** Unlike multiple cases described in the related work, de-identification of laboratory reports in STEM remained a non-trivial labor-intensive task. Applied out of the box, a machine learning toolkit and an artificial neural network technique did not enhance performance of the brute forth approach based on user list matching.
- **Directions for Future Research:** Customized tagging and training on the STEM corpus were presumed to advance outcomes of machine learning and predominantly artificial intelligence methods. Application of other natural language toolkits may lead to deducing a more effective solution.

Keywords: artificial intelligence, customized tagging, de-identification, machine learning, OpenNLP and NeuroNER toolkits, writing analytics, STEM Writing

1.0 Background

De-identification, anonymization, and pseudonymization are important processes in writing analytics. Various laws regulating privacy, use of personal information, and related issues continue to receive international attention. Multiple regulations demand data anonymization, pseudonymization, or the securing of personal information in cases where deidentification methods such as those described in this paper could (and perhaps should) be employed to protect human subjects. Completed as part of a larger research project for designing and launching artificial intelligence in the form of deep artificial neural network (DANN) models, the present study[1] describes how a corpus of laboratory reports in STEM ensured protection of human subjects by maintaining the confidentiality of data through corpus de-identification.

Specifically, a corpus of student lab reports in STEM was extracted from the MyReviewers (MyR) data warehouse (Rudniy, 2018) holding student corpora and supplementing information for certain courses taught at University of South Florida and several participating higher education institutions. MyR (Moxley, 2013; Moxley & Eubanks, 2016) is a sophisticated web application for writing projects in English and STEM fields, empowering students to submit assignments, receive feedback from peers and instructors, and comment on the work of other students. Reviews may be completed in several *modi operandi*, with rubric feedback being the most frequently employed. In this mode, a reviewer would designate textual feedback and numerical scores according to several criteria. Subsequently, MyR calculates an overall score as defined in the rubric and converts it to a letter grade. Additionally, except for rubric feedback, reviewers are advised to use Floating Comments, which can be added by highlighting text and adding a floating note. MyR also contains a library of Community Comments prepared for several fields of study using expert knowledge, most frequent comments, and electronic textbooks supplied with MyR.

Information collected with MyR and similar platforms (e.g., Eli Review, Write Lab) is extremely important for the field of writing studies. Multiple aspects were investigated and remain in focus of several research groups. Corpora produced with the use of MyR and its data warehouse were examined in a number of studies focusing on issues such as peer review (Moxley, 2017; Moxley et al., 2017; Ross et al., 2017), writing constructs assessment (Ross & LeGrand, 2017), corpus analysis (Anson & Anson, 2017; Aull, 2017; Leijen & Moxley, 2017; Moxley, 2017), writing program administration (Donahue et al., 2017; Kanuppinen et al., 2016; Moxley & Eubanks, 2016), STEM education (Moxley, 2016; Moxley, 2017; Donahue et al., 2017; Moxley et al., 2016), writing analytics (Elliot, et al., 2016; Moxley, 2017; Moxley & Walkup, 2016; Ross et al., 2016), and the role of instructor in peer feedback (Ross & LeGrand,

[1] The current work was completed as a constituent phase of NSF SBIR project # 1721749, which aimed to design deep artificial neural network (DANN) models generating automated feedback and scoring in order to help students improve their laboratory reports in STEM before submission. The work was completed under Aspire INV-A-021732 IRB approval. In addition, the planning work was completed under NSF Promoting Research and Innovation in Methodologies for Evaluation (PRIME) Program Award 1544239, which explored the role of instructor and peer feedback in improving STEM student writing.

2017; Anson & Anson, 2017; Moxley et al., 2016). Additionally, a number of research outcomes were produced under NSF Award 1544239: Collaborative Research: The Role of Instructor and Peer Feedback in Improving the Cognitive, Interpersonal, and Intrapersonal Competencies of Student Writers in STEM Courses.

1.1 Reasons for De-Identification

The STEM dataset contained information protected by FERPA (Family Educational Rights and Privacy Act), in particular, names of students who were either writing a report, completing an assignment as a member of a team, or facilitating the process in the role of a teaching assistant. In certain cases, students also included several digits of their numeric student IDs. De-identification of the latter was a straightforward task with pattern matching implemented with a computing technique known as regular expressions. Only the last four digits were preserved when a larger number was included in a text.

Along with de-identifying student names, a decision was made to wipe instructor information as well to preserve privacy and avoid possible negative consequences in the future. A growing trend of preserving human subjects' privacy exists in the United States and internationally. A detailed overview of the current state of affairs and a historical background are discussed in Section 2.0, Literature Review.

1.2 De-identification Approaches

It was noticed that in most cases names of students and instructors were located within the initial 400 characters of a text. The rest of the text might contain the name of the writing author when it was placed in pages' headers or footers, commonly accompanied by page numbers. On the other hand, the rest of the text commonly contained names of authors of referenced work. Such author names should not be removed, in order to preserve the information as devised by a student.

Advances demonstrated in the related work were not directly applicable to the current effort since not all personal names needed to be removed from texts—only those of students and instructors—while the rest of the names such as authors of related work must be preserved. STEM corpus was accompanied by supplementary data stored in the MyR data warehouse (DW) comprised of pseudonymized information—student and instructor IDs were replaced with artificial identifiers while preserving a lookup table in a separate secure location.

Following the course of the project, student writings were anonymized by matching student and instructor names stored in the lookup table and replacing matched names with the underscore "_" symbol. Not all the elements of the personal information were matched to the MyRerviewers users' list. In order to reduce manual processing, the research question hypothesized that an open-source natural language processing package or an artificial intelligence model for named entity recognition (NER) would improve the number of correct matches. The employment of NER tools was aimed at identifying and locating named entities of interest in texts. The recorded

information was used to subsequently replace the targeted NER types with the underscore symbol.

2.0 Literature Review

Relevant to the technical details of the project is its origin in protection of human subjects. Important to the study, therefore, is an analysis of laws and regulation relevant to privacy and security, where de-identification and pseudonymization may be applied for enforcement and compliance purposes.

The earliest act described in this section was ordained in 1946, and the latest General Data Protection Regulation (GDPR) enacted by the European Union in 2018 affected multiple companies and services with online presence. The literature review provides a description of personally identifiable information, pseudonymization, and de-identification, then turns to an overview of research studies on anonymity and de-identification. The review concludes with an outlook on applications in education and healthcare.

2.1 European Union Privacy Legislation

The European Union has a long history of protecting privacy and personal data. The effort began in 1980, when the Organization for Economic Cooperation and Development (OECD) developed Recommendations of the Council Concerning Guidelines Governing the Protection of Privacy and Trans-Border Flows of Personal Data (OECD Guidelines), which did not have the force of law. The Guidelines contained eight major principles for national data handling: (1) the Collection Limitation Principle demanded lawful data collection and appropriate notification or consent of data subjects, (2) the Data Quality Principle required the data to be relevant and necessary to the cause, and kept accurate, complete, and up-to-date, (3) the Purpose Specification Principle requested disclosure of goals at the time of collection, (4) the Use Limitation Principle prescribed not to share personal data except for the stated purposes except when required by the authority of law or with an additional consent of data subjects, (5) the Security Safeguards Principle commanded to reasonably secure the data, (6) the Openness Principle instructed to provide means to inform about developments, policies, and practices applied to the collected personal information, (7) the Individual Participation Principal asked to allow data subjects access to, communication about, and deletion of the collected personal data, and (8) the Accountability Principle specified that data subjects had to have a way to hold data collectors accountable for not following principles (1) through (7). The four principles of international application advised member countries of data processing, re-export, transborder data flows, and national legislation.

In 2013, revisions focusing on practical implementations of privacy protection by mitigating risks and interoperability improvement were introduced to the OECD Privacy Guidelines (OECD Privacy Guidelines, 2013). In particular, issues of data anonymization, anonymity, pseudonymity, and re-identification were addressed. An additional report accompanying the 2013 OECD revisions specified several issues, which were listed as suggested directions for

future work. Specifically, the Guidelines concluded that although anonymization and de-identification methods were capable of preserving privacy in data analytics, not all techniques were equally vigorous. The report questioned the role of anonymization and de-identification in settings when re-identification remained a persistent risk; the report also questioned whether a set of different identifiability degrees should be established and whether anonymization and other privacy-establishing methods were capable of establishing the balance between personal privacy and business use.

Except for the OECD Guidelines, the EU adopted the Data Protection Directive 95/46/EC in 1995, which regulated processing of personal data, addressing anonymization and de-identification broadly. It requested member states to provide personal data in a form permitting identification of subjects for the declared purposes and placed responsibility on member states for applying adequate processing to personal data when intended for longer historical, statistical, or scientific use (Directive 95/46/EC).

Superseding Directive 95/46/EC, the new data protection framework consisting of EU General Data Protection Regulation 2016/679 GDPR (Regulation EU 2016/679) and Directive EU 2016/680 took effect on May 25th, 2018, aimed at improving protection of natural persons (distinguished from business entities); prevention, investigation, detection, or prosecution of criminal offences; and free movement of such data by the means of governing personal data processing by competent authorities. GDPR stated that it did not apply to anonymous information when a data subject was no longer identifiable. Recital 78 of GDPR prescribed that data controllers apply measures and policies to comply with principles of data protection by design and by default, thus pseudonymizing personal data as soon as possible, allowing monitoring of data processing, and enabling creation and improvement of security features (Regulation EU 2016/679).

2.2 International Privacy Legislation

The OECD Privacy Guidelines were used by a number of countries as basis for national privacy protection practices, such as the 1988 Australian Privacy Act (Ludwig, 2009) with eleven information privacy principles and its 2001 amendment embracing identifiers, anonymity, and transborder data flows; the 1993 New Zealand Privacy Act (Power, 2008) which, in particular, elaborated on unique identifiers; the 2001 Canadian Personal Information Protection and Electronic Documents Act (Gilbert, 2009); the 2001 Korean Act on Promotion and Communications Network Utilization and Data Protection Act and its 2009 revisions urging the development of security measures for personal data and youth protection (Gilbert, 2009); the 2003 Japanese Act on the Protection of Personal Information on collection, use, and disclosure of personal information (Gilbert, 2009); the 2010 Mexican legal implementation of the OECD Guidelines (Decree for Federal Law), and the 2010 Turkish constitutional amendment for protection of personal data (Gilbert, 2009). The OECD Guidelines were also applied in the Privacy Framework for more than twenty countries of Asian-Pacific region participating in the

Asia-Pacific Economic Cooperation (APEC Privacy Framework, 2005). Gilbert (2009) describes a number of laws and policies related to personal data regulations passed by more than fifty nations from Argentina to Uruguay.

2.3 U.S. Legislation

The United States has a long history of handling personal information and privacy. One of the earliest mentions of personal privacy in U.S. laws appeared in the Administrative Procedure Act of 1946, which stated that public records preserved by government agencies shall be made available except cases with a good cause for confidentiality (Administrative Procedure Act of 1946). The law was reformed by the Freedom of Information Act (FOIA) of 1966, which regulated information disclosure by the U.S. government, prohibiting unwarranted invasion of personal privacy in personnel and medical files (Freedom of Information Act of 1966). It was amended with Electronic FOIA in 1996, which targeted electronic records.

The Fair Credit Reporting Act was passed in 1970 and reformed multiple times since then. The law recognized the need to ensure consumer information privacy, accuracy, and fairness within the data preserved by consumer reporting agencies (Fair Credit Reporting Act of 1970). Gellman (2017) has traced U.S. information privacy regulations and policies to 1973, when the Health, Education and Welfare (HEW) Advisory Committee on Automated Data Systems issued the Code of Fair Information Practices, which discussed safeguards, principles, and recommendation for personal data privacy (Records, Computers and the Rights of Citizens).

The Privacy Act of 1974 stated that the right to privacy was protected by the Constitution, that computers and information technology greatly magnified potential harm to personal privacy, and that it was necessary to regulate the collection, maintenance, use, and dissemination of federal agencies data. In this regard, the Act provided guidelines against invasion of personal privacy, record matching, data security, and destruction, among other issues (The Privacy Act of 1974).

Historically, other notable U.S. laws and regulations affecting privacy are important to the background of the present study. The Right to Financial Privacy Act of 1978 prescribed that government agencies must notify financial institutions' customers about accessing their records (Right to Financial Privacy Act of 1978). The Privacy Protection Act of 1980 protected journalists and newsrooms (Privacy Protection Act of 1980). The Cable Communications Policy Act of 1984 included a section on protecting subscriber privacy, regulating use of personally identifiable information (Public Law 98-549). The Electronic Communications Privacy Act (ECPA) of 1986 updated the Federal Wiretap Act of 1968. ECPA protected wire, oral, and e-communications in real time, in transit, and at rest. The act applied to phone, email, and electronic information, assigning varying levels of protection to different types of data (Electronic Communications Privacy Act of 1986). The CLOUD Act of 2018 may be considered as an update to ECPA of 1986, regulating government agencies' access to data stored overseas (H.R.4943).

The Video Privacy Protection Act of 1988 prevented disclosure of personally identifiable data on rental or sales records of video cassette tapes or similar audio-visual material (Video Privacy Protection Act of 1988). The Driver's Privacy Protection Act of 1994 prohibited the release or use of personal information collected by the departments of motor vehicles, with later amendments allowing data sharing after obtaining permissions from individuals. The Telephone Records and Privacy Protection Act of 2006 improved protection of the fraudulent acquisition or unauthorized disclosure of phone records (H.R. 4709). The Do-Not-Call Implementation Act of 2003 authorized the National Do Not Call Registry, which was made to establish compliance with the Telephone Consumer Protection Act of 199,1 restricting the use of telephone equipment and addressing privacy rights (H.R. 395, Telephone Consumer Protection Act).

Following the attacks on the U.S. on September 11, 2001, the USA PATRIOT Act of 2001 was a subject of criticism due to its provisions of electronic surveillance and invasion of privacy (USA PATRIOT Act of 2001). The Personal Data Privacy and Security Act of 2009 was aimed at prevention and mitigation of identity theft, privacy protection, commanding notifications of security breaches, and enforcing mishandling of personally identifiable data (S. 1490). The USA FREEDOM Act was viewed as legislation restoring privacy rights and ending bulk data collection by the government agencies (USA FREEDOM Act of 2015; Leahy, 2015.).

2.4 Personally Identifiable Information, Pseudonymization, and De-identification

Important to a discussion of student information is FERPA (Family Educational Rights and Privacy Act of 1974), which covered public and private elementary, secondary, or post-secondary schools or education agencies that received federal funding. Under FERPA, students were given the right to inspect or make corrections to their educational records or prohibit the release of personally identifiable information. Students were also given an option to receive a copy of their institution's policies on access to educational data. As well, FERPA forbade disclosing personally identifiable information without written consent. The act had important exemptions allowing release of personal data without student's or parent's consent to (1) school officials with a legitimate educational interest; (2) other institutions where a student sought or intended to enroll; (3) education officials for audit and evaluation purposes; (4) accrediting organizations; (5) parties in connection with financial aid to a student; (6) organizations conducting certain studies for or on behalf of a school; (7) comply with a judicial order or subpoena; (8) in the case of health and safety emergencies; and (9) state and local authorities within a juvenile justice system (Family Educational Rights and Privacy Act of 1974).

2.4.1 Personally identifiable information. In its GAO-08-536 Privacy Protection Alternatives report, the U.S. Government Accountability Office referred to personally identifiable information as any data about an individual including (1) any information that can be used to distinguish or trace one's identity, e.g., a name, a Social Security Number, date of birth, mother's maiden name, etc. and (2) other information, which can be linked to a person, e.g., medical, educational, financial, and employment information (GAO, 2008; Yoose, 2017).

2.4.2 Pseudonymization and pseudonymous data. GDPR described pseudonymization as the processing of personal data in a way that it cannot be linked to specific data subjects without separately-stored supplementary information. Contrary to pseudonymous data, anonymous information cannot be used to identify a natural person (Regulation EU 2016/679).

2.4.3 De-identification and de-identified information. O'Keefe and colleagues (2017) described de-identification as a process for removing or replacing direct identifiers, which may be followed by removing, making obscure, altering, or protecting data to prevent identification of an individual. The Privacy Act considered data to be de-identified when "the information was no longer about an identifiable individual or any individual who was reasonably identifiable."

2.5 Related Work on Anonymity and De-Identification

Sweeney (2002) designed a k-anonymity protection model and applied it to structured data, where n attributes referring to different persons would have k duplicates, with larger k values establishing higher degree of anonymity. Sweeney's anonymity model may be considered as an antipode to a candidate key in relational model. A data tuple was considered k-anonymous when surrounded with (k-1) tuples with the same values in n common attributes. Sweeney described several possible re-identification attacks and a theoretical background without specifying a method for automated de-identification. The same author showed, based on 1990 U.S. Census data, that 87% of the U.S. population were possibly identifiable from the combination of five-digit zip code, gender, and date of birth (Sweeney, 2000). In another work (Sweeney, 1998), the 1997 voting list of Cambridge, MA, was used to re-identify 29% of voters by birth dates and gender, 69% by birth date and five-digit zip code, and 97% of voters by birth date and full postal code. The study also presented a computer program, Datafly, for de-identifying structured data stored in a relational database.

Kumar and Helmy (2009) analyzed anonymity in wireless networks, where privacy can be compromised by deducing user identity from a combination of a MAC address—the identifier assigned to each network interface controller—and several other components of Wi-Fi log files, such as start time, duration, access point, etc. Kumar and Helmy described several attack scenarios and de-identification approaches, including k-anonymity (Sweeney, 2002) and l-diversity (Machanavajjhala et al., 2006).

Drachsler et al. (2010) discussed privacy concerns related to data re-identification applying Web 2.0 website information and provided suggestions for policies to be created to address these issues. Khalil and Ebner (2016) provided an overview of de-identification techniques applied to structured data and proposed applying a combination of hashing, suppression, masking, swapping, and noising for anonymization purposes.

To facilitate a general de-identification process, O'Keefe and colleagues (2017) laid out a decision-making framework overviewing legislation, privacy and ethics, de-identification, and the Five Safes framework, also known as a VML Security Model. O'Keefe and colleagues also established common options for data access and a de-identification framework.

2.6 Applications in Education

In learning analytics, tracking students' performance is needed to identify students at risk, interventions for correcting performance, forecasting, and other applications (Wachtler, 2016). It is well known that tracking student interactions in online social networks could reveal sensitive information on their identities (Boyd, 2008).

Big data technologies can be applied to student data for decision-making in instruction, student competencies analytics, predicting outcomes, monitoring at-risk students, and providing academic and career guidance. For additional insights, student data on enrollment and performance may be merged with a large number of additional variables aiming to determine placement into a particular course or university. With such data, multiple studies may be constructed: for evaluating instructors and whole institutions; for informed instructional design; and for improving pedagogical methodologies and instruction quality (Zeide, 2016).

The field of learning analytics identified two major dimensions of studies on student data, in reflection and prediction. Reflection was described as analytics for self-evaluation to obtain self-knowledge, monitor at-risk students, and suggest interventions. Prediction would model learners' performance and call for an early intervention or an adaptive curriculum offering additional high-complexity tasks for overachievers (Greller & Drachsler, 2012).

On the other hand, a number of concerns have been raised on student data security, mismanagement, and misuse by educational institutions and involved parties. Unauthorized access, unintentional disclosure, identity theft, publicizing sensitive information and using incorrect or outdated information for making decisions affecting students' future opportunities, discriminating against students based on their past performance, and repurposing student data to maximize profits by for-profit entities were among the hazardous aspects (Zeide, 2016).

Research studies on student writing, the topic of the present study, fall into the category of research on human subjects, which is susceptible to government regulations, such as The Family Educational Rights and Privacy Act. The Office for Human Research Protections described unanticipated problems which may arise during the course of research in the Guidance on Reviewing and Reporting Unanticipated Problems Involving Risks to Subjects or Others and Adverse Events. In many cases, approval of an Institutional Review Board (IRB) and additional ethics training were required for a researcher to begin a study. Relevant here is The Common Rule and its requirement that, when reviewing research proposals, IRB members must determine if adequate provisions for protecting the privacy of subjects and maintaining the confidentiality of data were made by an investigator (Kumar & Helmy, 2009; Machanavajjhala et al., 2006; Phelps-Hillen, 2017). The potential loss of confidentiality was among the red flags raised by IRBs, which could halt any project. Data de-identification was a common way of addressing this risk. This task required thorough processing since ineffectively anonymized data can be subsequently re-identified by using indirect markers and identifiers. We turn now to that task.

3.0 Research Question

Machine learning and artificial neural network-based techniques demonstrated significant advances in NER and de-identification for a number of tasks. Evaluation of these methods on the problem of student lab reports anonymization would facilitate an otherwise labor-intensive process. Due to the time constraints, a promising approach was to apply NER methods with pre-trained models, thus eliminating the time needed for additional tagging and training.

The research question of this paper was formulated as follows: Given the STEM dataset of laboratory reports, what was the most efficient method for de-identification of student and faculty entities among (1) search and replace using a lookup table of users' personal names, (2) application of a named entity recognition technique from a natural language processing toolkit based on machine learning, and (3) a deep artificial neural network for named entity recognition?

4.0 Research Methodology

4.1 De-identification with Users' List

A lookup table accompanying the MyR DW contained a list of first and last names linked to artificial numeric IDs. Each lab report had a corresponding record with supplementary information, including artificial IDs, which could be used to reference the actual name of a writer or grader. Several challenges brought unforeseen complications. For example, laboratory projects in STEM were frequently group assignments; however, only the person who prepared the report had to register in MyR. Since full class rosters were not provided by the university, the names of the report preparer's teammates were not available.

Additional challenges were brought by frequently misspelled graders' names, which made exact matching impossible. As well, writers included names of instructors or teaching assistants (TA) who were not on the users' list—some educators participated in teaching or grading but were never registered with MyR.

It was noticed that the names of students, TAs, and instructors were included in the beginning of a text within the initial 400 characters, while the rest of the laboratory report could contain its author's name either in page headers or footers. Nevertheless, their names had to be de-identified. To accomplish this, words comprising the initial 400 characters of a text were matched against the list of users in the same semester and in the same class as the text author. Manual inspection demonstrated unsatisfactory results with a number of names still present in texts.

To minimize manual correction of the unmatched personal information, words within the 400 characters in the beginning of each text were matched against the full list of MyR users. This approach improved the results, while at the same time increasing the number of false positives—words not comprising personal information were also affected (e.g., April may be the first name or the name of the month). Anonymization of instructors' and TAs' names was not a routing task as well since writers frequently used spelling varying from the users' list. The remaining issues

necessitated manual processing, which was also used to verify and potentially correct the rest of the personal data.

4.2 De-identification with a Named Entity Recognition Toolkit

Apache OpenNLP library was selected to conduct named entity recognition, which was listed among the toolkit capabilities along with other common NLP tasks, such as chunking, co-reference resolution, document categorization, language detection, lemmatization, parsing, part-of-speech tagging, sentence segmentation, and tokenization. While OpenNLP was capable of working with any language, trained models were provided for Danish, Dutch, English, German, Portuguese, Spanish, and Sami. Applications to other languages or custom entities detection required custom training and a tagged corpus, where the start and end of each new entity were tagged with angle-bracket markup (e.g., <START:entity> entity text <END>).

Natively, OpenNLP can be embedded into a custom Java program via its application programming interface or using command line instructions. R programming language and a package with an interface to OpenNLP (2016) were used for programming implementation in this work. The OpenNLP toolkit was chosen due to its high performance—it achieved 0.94 precision, 0.75 recall, and 0.83 *F*-measure on the MUC-7 data (MUC-7 dataset). In a study evaluating NLP toolkits, Pinto, Oliveira, and Alves (2016) found that OpenNLP demonstrated the best overall performance for a corpus of formal text. Additionally, the Apache Software License allowed commercial use, facilitating seamless future integration into another program, unlike several other named-entity recognition packages covered with research-only licenses or a General Public License (Ingersoll et al., 2013).

Pre-trained English models supplied with OpenNLP were used to avoid additional tagging and training. When looking for names, OpenNLP started by splitting text into sentences. The sentences were then tokenized, and the list of names was returned. While processing a text, the package kept a record of previously recognized names to apply to subsequent occurrences of the same words. Normally, sentences were considered separately to prevent the program from identifying a name crossing sentence boundaries. For identified names, their first and last character positions were recorded, allowing for the determination of locations in the original text.

4.3 De-identification with an Artificial Neural Network

Since the seminal work by McCulogh and Pitts (1943), DANNs evolved significantly, overcoming a limited learning ability of single-layer perceptions (Minsky & Rapert, 1969) by adapting multi-layer models (LeCun, 1986; Parker, 1985) and finally emerging to complex DANN architectures providing state-of-the-art results in the image recognition and NLP domains (Collobert et al., 2011; Gulshan et al., 2016; Hinton et al., 2012; Kalchbrenner, 2014). DANNs outperformed a number of machine learning algorithms when used for named entity recognition (Collobert et al., 2011; Dernoncourt et al., 2017; Dernoncourt et al., 2016; Labeau, et al., 2015; Lample et al., 2016; Lee et al., 2016). DANNs were capable of learning NLP features jointly

with model parameters from training data, which explained their superior performance (Dernoncourt, et al., 2017).

The NeuroNER open-source system achieved near state-of-the-art results on the i2b2 2014 (Sang & De Meulder, 2003) and CoNLL 2003 (Stubbs et al., 2015) data gaining respectively 0.905 and 0.977 values in *F*1-measure (a measure that considers both precision and recall). NeuroNER relied on TensorFlow, a machine learning system developed at Google Research and capable of distributed computation on a large number of machines or graphics processing units (GPU) cards, commonly used for ANN training and evaluation (Abadi et al., 2016).

The NeuroNER system was selected due to its demonstrated performance, flexibility, and adaptability. As OpenNLP, NeuroNER eliminates the time-consuming corpus labeling phase by using out-of-the-box pre-trained models such as word2vec (Mikolov et al., 2013a; Mikolov et al., 2013b; Mikolov, et al., 2013s) or GloVe (Pennington et al.,, 2014) word embeddings. NeuroNER consisted of three layers. A character-enhanced token-embedding layer mapped tokens to their vector representations, which were passed to the label prediction layer producing probabilities of a NER label for a vector, and a label sequence optimization layer made final label assignments. (Dernoncourt et al., 2017)

The system allowed for the modification of algorithm parameters within a configuration file, adjusting the number of central processing unit (CPU) threads, the number of GPUs, dimensions of Long Short-Term Memory neural network embeddings, character-based token embeddings, dropout rate, maximum number of epochs, and other parameters. Additionally, NeuroNER allows integration with BRAT, a web-based corpus annotation and visualization tool, allowing researchers to tag various corpora with its user-friendly interface, store produced markup in BRAT format, or convert the data to external markup formats (Stenetorp et al., 2012).

4.4 Experiment Design

De-identification with the users' list was implemented with a custom program designed in Microsoft Visual Studio C# programming language and multi-threading for parallel processing. The OpenNLP NER toolkit was applied using RStudio and an R library providing an interface to the original OpenNLP code in Java. Both C# and R programs were executed in the Windows Server environment in the Microsoft Azure cloud. A NeuroNER model was executed in Ubuntu Linux Server command line with a number of parameters overriding the default settings stored in the configuration file. The NC-24 Microsoft Azure virtual machine with four GPUs was employed in the experiment.

As mentioned above, NER methods were applied with pre-trained models, thus avoiding design of a training set. A test set consisted of 1,000 lab reports produced by University of South Florida students in CHM 2045 General Chemistry I and CHM 2046 General Chemistry II, randomly selected from a bigger corpus used in NSF SBIR project # 1721749. An excerpt from a sample de-identified text is depicted in Figure 1.

5.0 Results

The output of conducted experiments was stored in text files, which were imported into a relational database, where SQL queries were used for filtering and managing data. The resulting data consisted of texts split into tokens, where each token was assigned an NER tag. NeuroNER used Beginning-Inner-Outer (BIO) tagging format. In particular, B-PER denoted the beginning of a person entity, I-PER for an inner token continued a person entity, and O tag was used for all outer tokens not related to persons' names detected by the system (Jiang, 2012; Troyano et al., 2004).

Confusion matrices (shown in Table 1, Table 2, and Table 3) precision, recall, and *F*1 measures (shown in Table 4) were calculated as evaluation measures following a common approach in the machine learning and NER fields. A confusion matrix for a single-class problem consisted of four quadrants: true negatives (TN), false negatives (FN), false positives (FP), and true positives (TP). A confusion matrix can be used to calculate a number of other performance measures used in machine learning, such as precision, recall, and *F*1.

Table 1

Users' List Confusion Matrix

Users' List	Predicted **Not** a Student/TA/Instructor	Predicted Student/TA/Instructor
Actual **Not** a Student/TA/Instructor	1,749,371 *True Negatives (TN)*	598 *False Positives (FP)*
Actual Student/TA/Instructor	723 *False Negatives (FN)*	2,200 *True Positives (TP)*

Table 2

OpenNLP Confusion Matrix

OpenNLP	Predicted **Not** a Student/TA/Instructor	Predicted Student/TA/Instructor
Actual **Not** a Student/TA/Instructor	1,743,502 *True Negatives (TN)*	6,467 *False Positives (FP)*
Actual Student/TA/Instructor	2,504 *False Negatives (FN)*	419 *True Positives (TP)*

Table 3

NeuroNER Confusion Matrix

NeuroNER	Predicted **Not** a Student/TA/Instructor	Predicted Student/TA/Instructor
Actual **Not** a Student/TA/Instructor	1,740,325 *True Negatives (TN)*	9,644 *False Positives (FP)*
Actual Student/TA/Instructor	1,298 *False Negatives (FN)*	1,625 *True Positives (TP)*

In the current work, true negatives denoted those tokens found in text that in fact were not names of a student, instructor, or a TA and were truly predicted by an algorithm as such. False negatives were those tokens that in fact were names of a student, instructor, or a TA, but were falsely predicted by an algorithm as not. True positives were those tokens that in fact were students', instructors', or TA's names and were correctly predicted by an algorithm. False positives were those tokens that in fact were not students', instructors' or TA's names, although predicted by an algorithm as such.

Visual comparison of Tables 1 - 3 shows that the number of true positives was the highest for the user's list method denoting its best performance, NeuroNER was second best, and the OpenNLP numbers were the lowest. On the other hand, the number of incorrectly labeled student, instructor, or TA names, or the number of false negatives was the lowest for the user's list technique, showing its superior performance, which was followed by NeuroNER with the second-best number, and OpenNLP concluding the list.

Precision *P*, recall *R*, and their harmonic mean *F*1 were calculated as shown in Table 4 using the confusion matrix numbers to compare performance of the three evaluated techniques.

Table 4

Precision, Recall, and F*1*

	Precision *P*	Recall *R*	Harmonic Mean *F*1
Formula	$P = \frac{TP}{TP + FP}$	$R = \frac{TP}{TP + FN}$	$F1 = 2\frac{P * R}{P + R}$
Users' List	**0.79**	**0.75**	**0.77**
OpenNLP	0.06	0.14	0.09
NeuroNER	0.14	0.56	0.23

In this work, precision *P* was the capability of a method to correctly mark tokens as names of students, instructors, or TAs. Recall *R* showed the ability of a technique to correctly identify *all* names of instructors, students, or TAs. *F*1 was a harmonic mean combining both precision *P* and recall *R* within a single metric. Precision, recall, and *F*1 ranged from 0 (or 0%) being the worst value to 1 (or 100%), which is the best possible value.

Table 4 demonstrates that the users' list method significantly outperformed both NER toolkits when applied to the specific de-identification task, with *P*=0.79, *R*=0.75, and *F*1=0.77. The artificial neural network-based toolkit NeuroNER gained a relatively high value of recall *R* of 0.56. Overall, the evaluation of the three selected methods demonstrated that automating de-identification was not possible in the setting of this work. For preserving privacy, 100% recall must be achieved. Nonetheless, applying the users' list method with subsequent manual processing would allow for reaching the targeted recall value and significantly reduce manual labor.

6.0 Discussion

The task for locating persons' names belonging to student writers, TAs, or instructors while omitting other names appearing in texts was not trivial. This factor impacted performance of OpenNLP and NeuroNER, which otherwise were significantly higher as described in Section 4.0, Research Methodology. In this work, NER packages were applied out of the box with the included pre-trained models, which eliminated time-consuming manual tagging of the corpus and the subsequent training phase.

The demonstrated low numbers of precision *P*, recall *R*, and *F*1 measures encouraged use of custom models, trained on the same STEM corpus as used in this work. Thus, OpenNLP and NeuroNER would infer statistical properties of the corpus, subsequently improving performance. It is worth noting that the goal of applying the machine learning and ANN packages was not to fully automate the process, but to reduce the amount of manual labor that would still be required for verification and validation purposes. In this work, when actual de-identification took place, student, TA, and instructor names were substituted with the underscore symbol as shown in Figure 1.

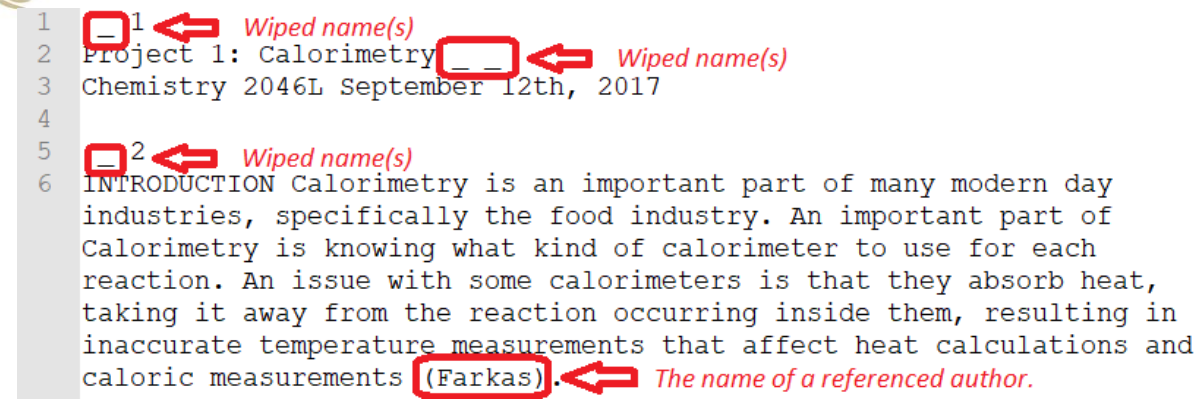

Figure 1. Sample de-identified text.

Thus, when a de-identification error occurred and an appropriate name was not wiped out from the text, it would become clear to the reader that de-identification failed and re-identification was possible. To illustrate this, consider line 2 in Figure 1, saying instead of "Project 1: Calorimetry _ _" as the correct de-identification process would produce, another line with "Project 1: Calorimetry _ Fitzgerald." To avoid such issues and to mask possible de-identification failures, it would be appropriate instead of substituting a name with an underscore symbol, to substitute a name with another name from publicly available lists. Such a replacement would either reduce the chances of subsequent re-identification or make it impossible.

It is worth noting that the approaches demonstrated in this work are applicable to other datasets in the domain of writing analytics. OpenNLP and NeuroNER may be applied to other corpora. Further adaptation may be done through designing custom training sets by tagging existing or introducing new named entities. Such effort requires significant manual processing, which can be accelerated with corpus markup and visualization tools.

7.0 Conclusions

As has been true for decades in multiple areas, especially in education, de-identification, anonymization, and pseudonymization remain important research issues. Various laws regulating privacy, use of personal information, and related issues were passed in several countries with the latest General Data Protection Regulation implemented on May 25, 2018. Multiple regulations assumed data anonymization, pseudonymization, or the securing of personal information, where automated de-identification and NER methods such as those described in this paper, could and should be employed.

Unlike cases described in the Literature Review section, de-identification of laboratory reports in STEM is a non-trivial labor-intensive task. The current work demonstrated that automation with the help of one machine learning and one artificial neural network technique did not improve results of the brute forth approach employing the list of users' names. This was caused by the complexity of the task, requiring removal of only particular names belonging to student writers, instructors, or teaching assistants.

The NER task described in this work is daunting since names of authors in cited sources and reference lists must be preserved, while personal names of writers, graders, and faculty must be removed. As illustrated in Figure 1, Farkas in line 6—a referenced author name—should be kept while a writer's name should be wiped from lines 1, 2, and 5. This example demonstrated a common case, when a writer's name was included in a page header with a page number. Thus, wiping out all the names in the beginning of texts would leave those appearing in the remainder. Machine learning NLP toolkits such as OpenNLP, DANN NLP systems, and NeuroNER are known to learn statistical dependencies from context. These novel methods are known for broader generalizability as described in Sections 4.2 and 4.3 of this paper. Thus, custom models trained to distinguish referenced authors from academic personnel may lead to better precision and recall.

8.0 Directions for Future Research

Several issues were assigned as directions for future work since they were out of scope of the current study. First, it would be of interest to train both OpenNLP and NeuroNER on the STEM corpus used in this project. For this purpose, the corpus must be manually tagged and split into train and test sets.

Second, OpenNLP and NeuroNER allow for adding new entity types, such as phone numbers, street addresses, and so on (Ingersoll, Morton, & Farris, 2013; Adding a New Entity Type, 2017). In this work, the B-PER and I-PER tags were used to denote students', instructors', and TAs' names. Dedicating an additional tag for the subset of personal names and using it for tagging a corpus may lead to performance improvement and should be investigated.

Third, we hypothesize that an artificial neural network given an appropriate training set and valid parameters will be capable of learning the difference between names of students or graders, which should be preserved, and other personal names and information. An extensive evaluation adjusting embedding size, type of embeddings, and other parameters is required to further study this issue.

Fourth, distinguishing writers' and graders' names from citations may not be necessary for certain NLP tasks such as performing analysis of context features of student writings (e.g., sentence structure or key term analysis). Thus, NER algorithms will be aimed at identification and removal of all personal names, potentially improving the outcomes.

Finally, it would be of interest to apply other NLP toolkits with NER functionality, which demonstrated top results, e.g., Stanford Core NLP, which is open-source and distributed under the General Public License (The Stanford Natural Language Processing Group).

Author Biography

Alex Rudniy is Assistant Professor of Computer Science at the University of Scranton. He has taught courses in data mining, programming, and massive data analysis at New Jersey Institute of Technology and at Farleigh Dickinson University. From 2015 to 2017, he served as Co-Principal

Investigator on NSF PRIME Award 1544239: Collaborative Research: The Role of Instructor and Peer Feedback in Improving the Cognitive, Interpersonal, and Intrapersonal Competencies of Student Writers in STEM Courses. In 2018, he served as Chief Technology Officer on NSF SBIR Award 1721749: SBIR Phase I: Artificial Intelligence, Scientific Reasoning, and Formative Feedback: Structuring Success for STEM Students. His research has appeared in *Assessing Writing*, *BMC Bioinformatics*, *Knowledge and Information Systems*, *International Journal of ePortfolio*, *Journal of Writing Assessment*, and *Research in the Teaching of English*.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., … Devin, M. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. Retrieved from https://arxiv.org/abs/1603.04467.

Adding a new entity type [Online forum comment]. (2017, 23 October). Retrieved from https://github.com/Franck-Dernoncourt/NeuroNER/issues/73.

Administrative Procedure Act of 1946. Retrieved from https://www.justice.gov/sites/default/files/jmd/legacy/2014/05/01/act-pl79-404.pdf.

Anson, I.G., & Anson, C.M. (2017). Assessing peer and instructor response to writing: A corpus analysis from an expert survey. *Assessing Writing, 33*, 12–24.

Apache OpenNLP Developer Documentation (2017). Retrieved from https://opennlp.apache.org/docs/1.8.4/manual/opennlp.html#intro.description

APEC Privacy Framework. (2005). Retrieved from https://www.apec.org/-/media/APEC/Publications/2005/12/APEC-Privacy-Framework/05_ecsg_privacyframewk.pdf.

Aull, L. (2017). Corpus analysis of argumentative versus explanatory discourse in writing task genres. *Journal of Writing Analytics, 1*, 1–47.

Bayardo, R.J., & Agrawal, R. (2005). Data privacy through optimal k-anonymization. *Proceedings: IEEE 21st International Conference on Data Engineering*, 217–228. Retrieved from https://doi.org/10.1109/ICDE.2005.42.

Beckwith, B.A., Mahaadevan, R., Balis, U.J., & Kuo, F. (2006). Development and evaluation of an open source software tool for deidentification of pathology reports. *BMC Medical Informatics and Decision Making, 6*(12), 1–10.

Boyd, D. (2008). Facebook's privacy trainwreck: Exposure, invasion, and social convergence. *Convergence: The International Journal of Research into New Media Technologies*, *14*(1), 13–20. Retrieved from http://dx.doi.org/10.1177/1354856507084416.

Carafe [Computer software] (2005). Available from https://sourceforge.net/projects/carafe/

Children's Online Privacy Protection Act of 1998 (COPPA). Retrieved from https://www.epic.org/privacy/kids/.

Coalition Letter against DOJ's XBD Bill. Retrieved from https://www.eff.org/document/2017-09-20-coalition-letter-against-dojs-xbd-bill.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research, 12*, 2493–2537.

Daries, J.P., Reich, J., Waldo, J., Young, E.M., Whittinghill, J., Ho, A.D., … Chuang, I. (2014). Quality social science research and the privacy of human subjects require trust. *Communications of the ACM, 57*(9), 56–63.

Decree for Federal Law of Protection of Personal Data in Possession of Individuals. Retrieved from http://www.dof.gob.mx/nota_detalle.php?codigo=5150631&fecha=05/07/2010.

Dehghan, A., Kovacevicr, A., Karystianisab, G., Keanead, J.A., & Nenadic, G. (2015). Combining knowledge- and data-driven methods for de-identification of clinical narratives. *Journal of Biomedical Informatics*, *58*(5), 53 –59.

Dernoncourt, F., Lee, J. Y., & Szolovits, P. (2017). NeuroNER: An easy-to-use program for named-entity recognition based on neural networks. Retrieved form https://arxiv.org/abs/1705.05487.

Dernoncourt, F., Lee, J. Y., Uzuner, O., & Szolovits, P. (2016). De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association. 24*(3), 596 – 606. doi: 10.1093/jamia/ocw156.

Directive (EU) 2016/680 of the European Parliament and of the Council. Retrieved from http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016L0680&from=EN.

Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. (1995). Retrieved from https://eur-lex.europa.eu/eli/dir/1995/46/oj.

Donahue, C., Elliot, N., Ross, V., & Moxley, J. (2017, January). *At the intersection of writing program administration, the digital humanities, and STEM education: Corpus methods as a lens for reader response*. Paper presented at the Modern Language Association Convention. Philadelphia, PA.

Douglass, M., Clifford, G., Reisner, A., Moody, G., & Mark, R. (2005). Computer-assisted de-identification of free text in the MIMIC II database. *Computational Cardiology*, *32*, 331–334.

Electronic Communications Privacy Act of 1986 (ECPA), 18 U.S.C. § 2510-22. Retrieved from https://it.ojp.gov/PrivacyLiberty/authorities/statutes/1285.

Elliot, N., Walkup, K., & Moxley, J. (2016). Preface to workshop two: Writing analytics, data mining, and writing studies. *Proceedings of the 9th International Conference on Education Data Mining*. Raleigh, NC: EDM.

Fair Credit Reporting Act. 15 U.S.C. §1618. Retrieved from https://www.ftc.gov/system/files/fcra_2016.pdf.

Family Educational Rights and Privacy Act of 1974 (FERPA). Retrieved from https://epic.org/privacy/student/ferpa/.

Federal Policy for the Protection of Human Subjects ('Common Rule'). Retrieved from https://www.hhs.gov/ohrp/regulations-and-policy/regulations/ OECD Privacy Guidelines (2013) common-rule/index.html.

Federal Trade Commission Enforcement of the U.S.-EU and U.S.-Swiss Safe Harbor Frameworks. Retrieved from https://www.ftc.gov/tips-advice/business-center/guidance/federal-trade-commission-enforcement-us-eu-us-swiss-safe-harbor.

Foufi, V., Gaudet-Blavignac, C., Chevrier, R., & Lovis, C. (2017). De-identification of medical narrative data. In R. Engelbrecht, R. Balicer, & M. Hercigonja-Szekeres (Eds.), *The practice of patient centered cure* (pp. 23–27). Amsterdam: IOS Press.

Freedom of Information Act of 1966. Retrieved from http://congressionaldata.org/the-original-text-of-the-freedom-of-information-act/.

GAO. (2008). *Privacy: alternatives exist for enhancing protection of personally identifiable information: Report to congressional requesters* (Report # GAO-08-536). Washington, DC: US Govt. Accountability Office. Retrieved from http://purl.access.gpo.gov/GPO/LPS111810.

Gellman, R. (2017). Fair information practices: A basic history. Version 2.18. Retrieved from https://bobgellman.com/rg-docs/rg-FIPshistory.pdf.

Gilbert, F. (2009). *Global privacy and security law*. Austin, TX: Wolters Kluwer Law & Business.

Greller, W., & Drachsler, H. (2012). Translating learning into numbers: A generic framework for learning analytics. *Educational Technology & Society*, *15*(3), 42–57.

Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Retrieved from https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#protected.

Gulshan V., Peng L., Coram M., Stumpe M.C., Wu D., Narayanaswamy A., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, *316*(22), 2402–2410.

Gupta, D., Saul, M., & Gilbertson, J. (2004). Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. *American Journal of Clinical Pathology*, *121*(2), 176–186. Retrieved from https://doi.org/10.1309/E6K33GBPE5C27FYU.

H.R. 395 — 108th Congress: Do-Not-Call Implementation Act. Retrieved from https://www.govtrack.us/congress/bills/108/hr395.

H.R. 4709 — 109th Congress: Telephone Records and Privacy Protection Act of 2006. Retrieved from https://www.govtrack.us/congress/bills/109/hr4709.

H.R.3103 - Health Insurance Portability and Accountability Act of 1996. Retrieved from https://www.congress.gov/bill/104th-congress/house-bill/3103/text.

H.R.493 - Genetic Information Nondiscrimination Act of 2008. Retrieved from https://www.congress.gov/bill/110th-congress/house-bill/493/text.

H.R.4943 - CLOUD Act. Retrieved from https://www.congress.gov/bill/115th-congress/house-bill/4943/text.

Haber, S., Hatano, Y., Honda, Y., Horne, W., Miyazaki, K., Sander, T., … Yao, D. (2007). *Efficient signature schemes supporting redaction, pseudonymization, and data deidentification* (Report # HPL-2007-191). Retrieved from http://hpl.hp.com/techreports/2007/HPL-2007-191.pdf.

Hay, M., Miklau, G., Jensen, D., Towsley, D., & Weis, P. (2008). Resisting structural re-identification in anonymized social networks. *Proceedings of the VLDB Endowment*, *1*(1), 102–114.

Health Insurance Portability and Accountability Act of 1996. Retrieved from https://www.gpo.gov/fdsys/pkg/PLAW-104publ191/pdf/PLAW-104publ191.pdf

Hinton, G. E., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.*, *29*(6), 82–97.

Ingersoll, G.S., Morton, T.S., & Farris, A.L. (2013). *Taming text. How to find, organize, and manipulate it*. New York: Manning Publications.

Jiang, J. (2012). Information extraction from text. In C.C. Aggarwal & C. Zhai (Eds.), *Mining text data* (pp. 11–41). New York: Springer Verlag.

Jolly, I. (2017). Data protection in the United States: Overview. Retrieved from https://uk.practicallaw.thomsonreuters.com/6-502-0467?transitionType=Default&contextData=(sc.Default)&firstPage=true&bhcp=1

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. Retrieved from https://arxiv.org/abs/1404.2188

Kanuppinen, A., Leijen, D., Moxley, J., & Wärnsby, A. (2016). *Pre-conference workshop: Responsible Action: International Higher Education Writing Research Exchange*. Presented at the Conference on College Composition and Communication Convention, Houston, TX.

Khalil, M., & Ebner, M. (2016). De-identification in learning analytics. *Journal of Learning Analytics*, *3*(1), 129–138. Retrieved from https://doi.org/10.18608/jla.2016.31.8.

Kudo, T., & Matsumoto, Y. A. (2004). Boosting algorithm for classification of semi-structured text. *Proceedings:Empirical Methods in Natural Language Processing 2004*, 301–308.

Kumar, U., & Helmy, A. (2009). Human behavior and challenges of anonymizing WLAN traces. *Proceedings from IEEE Globecom'09: Global Communications Conference*.

Labeau, M., L'oser, K., & Allauzen, A. (2015). Non-lexical neural architecture for fine-grained POS tagging. *Proceedings: 2015 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Lisbon, Portugal, 232–237.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. Retrieved from https://arxiv.org/abs/1603.01360.

Leahy, P. (2015). Bipartisan coalition led by Senators Lee and Leahy introduce legislation to ban bulk collection under Section 215 USA FREEDOM Act of 2015 would bring historic reforms to surveillance authorities. Retrieved from https://www.leahy.senate.gov/press/bipartisan-coalition-led-by-senators-lee-and-leahy-introduce-legislation_to-ban-bulk-collection-under-section-215.

LeCun, Y. (1986). Learning processes in an asymmetric threshold network. In E. Bienenstock, F, Fogelman-Soulié, & G. Weisbuch (Eds,), *Disordered systems and biological organizations* (pp. 233–240). Les Houches, France: Springer.

Lee, J.Y., Dernoncourt, F., Uzuner, O., & Szolovits, P. (2016). Feature-augmented neural networks for patient note de-identification. *Proceedings of the Clinical Natural Language Processing Workshop*, 17–22, Osaka, Japan.

Leijen, D., & Moxley, J. (2017, June). *The value of peer review across different institutional, national, and curricular contexts*. Paper presented at the 9th Conference of the European Association for Teaching Academic Writing (EATAW). University of London, London, UK.

Li, M. (2018). *Scalable natural language de-identification based on machine learning approaches* (Doctoral dissertation). Retrieved from https://etd.library.vanderbilt.edu/available/etd-03262018-113355/unrestricted/Li.pdf.

Lopez-Otero, P., Docio-Fernandez, L., Abad, A., & Garcia-Mateo, C. (August, 2017). *Depression detection using automatic transcriptions of de-identified speech*. Paper presented at INTERSPEECH 2017, Stockholm, Sweden.

Ludwig, J. (2009). *Australian government. Enhancing national privacy protection. Australian government first stage response to the Australian law reform commission* (Report # 108). Retrieved from https://www.alrc.gov.au/sites/default/files/pdfs/government_1st_stage_response.pdf.

Machanavajjhala, A., Gehrke, J., & Kifer, D. (2006). l-diversity: Privacy beyond k-anonymity. *Proceedings from ICDE'06: 22nd International Conference on Data Engineering*, 24–24.

McCulloch, W.S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*(4):115–133.

Meldau, E. (2018). Deep neural networks for inverse de-identification of medical case narratives in reports of suspected adverse drug reactions. Retrieved from https://kth.diva-portal.org/smash/get/diva2:1185934/FULLTEXT01.pdf.

Menikoff, J., Kaneshiro, J., & Pritchard, I. (2017). The Common Rule, Updated. *New England Journal of Medicine*, *82*, 613–615.

Meystre, S.M., Friedlin, F.J., South, B.R., Shen, S., & 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*(70), 1–16. Retrieved from https://doi.org/10.1186/1471-2288-10-70

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. Retrieved from https://arxiv.org/abs/1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Proceedings: Advances in neural information processing systems (NIPS 2013)*, 3111–3119.

Mikolov, T., Yih, W., & Zweig, G. (2013c). Linguistic regularities in continuous space word representations. *Proceedings: 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751.

Minsky, M., & Papert, S. (1969). *Perceptrons. An introduction to computational geometry*. Cambridge, Mass: M.I.T. Press.

Modifications to the HIPAA Privacy, Security, Enforcement, and Breach Notification Rules under the Health Information Technology for Economic and Clinical Health Act and the Genetic Information Nondiscrimination Act; Other modifications to the HIPAA Rules. Retrieved from https://www.federalregister.gov/documents/2013/01/25/2013-01073/modifications-to-the-hipaa-privacy-security-enforcement-and-breach-notification-rules-under-the.

Moxley, J. (2013). Big data, learning analytics, and social assessment. *The Journal of Writing Assessment*, *6*(1). Retrieved from http://www.journalofwritingassessment.org/article.php?article=68.

Moxley, J. (2016). *From FYC to professional & technical communication, partnerships with other U.S. post-secondary programs, NSF funding, and STEM education: Rethinking document critique, collaboration, and research across cultures and contexts*. Paper presented at the Computers and Writing, Rochester, NY.

Moxley, J. (2017a, January). *e-Portfolios and digital learning: The future of corpus studies in the domain of writing analytics*. Paper presented at the 8th Annual Forum on Digital Learning and ePortfolios, San Francisco, CA.

Moxley, J. (2017b, January). *Welcome to writing analytics*. Paper presented at the 4th International Conference on Writing Analytics: Writing Analytics, Data Mining, and Student Success 2017, St. Petersburg, FL.

Moxley, J. (2017c, March.) *Evidence-based framework for structuring learning opportunities: Peer review, STEM, and digital feedback's new culture*. Poster presented at the Carnegie Foundation Summit, San Francisco, CA.

Moxley, J., & Eubanks, D. (2016). On keeping score: Instructors' vs. students' rubric ratings of 46,689 essays. *Journal of the Council of Writing Program Administrators*, *39*(2), 53–80.

Moxley, J., Ross, V., Elliot, N., Rudniy, A., & Trauth, E. (2016). *The role of instructor and peer feedback in improving the cognitive, interpersonal, and intrapersonal competencies of student writers in STEM courses*. Presented at the International Writing Across the Curriculum Conference, University of Michigan, MI.

Moxley, J., Ross, V., & Trauth, E. (2016). *Making friends in STEM: Experiences, challenges, and triumphs in collaboration, data mining, peer review, and assessment*. Presented at the Council for Writing Program Administrators, Raleigh, NC.

Moxley, J., & Walkup, K. (2016). Mapping writing analytics. In *Proceedings of the 9th International Conference on Educational Data Mining*. Raleigh, NC: EDM.

Moxley, J., Wärnsby, A., Kauppinen, A., Leijen, D., Aull, L., Anderson, L., & Walkup, K. (2017, February). *Politeness, social and intrapersonal presence in student peer reviews: A cross-cultural analysis*. Roundtable held at Writing Research Across Borders IV, Bogata, Columbia.

MUC-7 dataset. Available from https://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_toc.html.

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. *Proceedings*: *2008 IEEE Symposium on Security and Privacy*, 111–125.

Neamatullah, I., Douglass, M. M., Lehman, L. H., Reisner, A., Villarroel, M., Long, W. J., … Clifford, G.D. (2008). Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, *8*(32), 1–17. Retrieved from https://doi.org/10.1186/1472-6947-8-32.

O'Keefe, C. M., Otorepec, S., Elliot, M., Mackey, E., & O'Hara, K. (2017). The de-identification decision-making framework, 1–76. Retrieved from: https://publications.csiro.au/rpr/download?pid=csiro:EP173122&dsid=DS2.

OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data. Retrieved from http://www.oecd.org/internet/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm.

OECD Privacy Guidelines (2013). Retrieved from: http://www.oecd.org/internet/ieconomy/privacy-guidelines.htm.

OECD. (2013). Privacy Expert Group Report on the Review of the 1980 OECD Privacy Guidelines, OECD Digital Economy Papers, No. 229, Paris: OECD Publishing. Retrieved from http://dx.doi.org/10.1787/5k3xz5zmj2mx-en.

Package 'openNLP'. (2016). Retrieved from https://cran.r-project.org/web/packages/openNLP/openNLP.pdf.

Parker, D. B. (1985). *Learning-logic: Casting the cortex of the human brain in silicon*. Technical Report Tr-47, Center for Computational Research in Economics and Management Science. Cambridge, MA: MIT.

Pennington, J., Socher, R., & C. Manning (2014). GloVe: Global vectors for word representation. *Proceedings: EMNLP 2014 Conference on Empirical Methods in Natural Language Processing*, 1532–1543.

Phelps-Hillen, J. (2017). *Institutional Review Boards and Writing Studies research: A justice-oriented study* (Doctoral dissertation). Retrieved from Scholar Commons Graduate Theses and Dissertations. http://scholarcommons.usf.edu/etd/6742.

Pinto, A., Oliveira, H.G., & Alves, A.O. (2016). Comparing the performance of different NLP toolkits in formal and social media text. *Proceedings: 5th Symposium on Languages, Applications and Technologies (SLATE 2016)*, 3:1-3:16. DOI 10.4230/OASIcs.SLATE.2016.3.

Power, S. (2008). Privacy (Cross-border Information) Amendment Bill. Government Bill. Retrieved from http://www.legislation.govt.nz/bill/government/2008/0221/latest/whole.html.

Privacy Protection Act of 1980. Retrieved from https://www.justice.gov/usam/criminal-resource-manual-661-privacy-protection-act-1980.

Public Law 111–5. Title XIII—Health Information Technology. Retrieved from https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/hitechact.pdf.

Public Law 98-549. Retrieved from https://transition.fcc.gov/Bureaus/OSEC/library/legislative_histories/1286.pdf.

Quinlan, J.R. (1993). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.

Records, computers and the rights of citizens. Report of the Secretary's Advisory Committee on Automated Personal Data Systems. (1973). https://epic.org/privacy/hew1973report/default.html

Regulation (EU) 2016/679 of the European Parliament and of the Council. Retrieved from: http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN.

Right to Financial Privacy Act of 1978. Chapter 35—Right to Financial Privacy. Retrieved from http://uscode.house.gov/view.xhtml?path=/prelim@title12/chapter35&edition=prelim.

Ross, V., Elliot, N., Rudniy, A., & Moxley, J. (2016). *Writing analytics, data mining, & writing studies*. Presented at the 9th International Conference on Educational Data Mining (EDM 2016), Raleigh, NC.

Ross, V., & LeGrand, R. (2017) Assessing writing constructs: Toward an expanded view of inter-rater reliability. *The Journal of Writing Analytics*, *1*, 227–275.

Ross, V., Liberman, M., Ngo, L., & LeGrand, R. (2017). Weighted log-odds-ratio, informative Dirichlet prior method to compare peer review feedback for top and bottom quartile college students in a first-year writing program. Retrieved from http://ceur-ws.org/Vol-1633/ws2-paper4.pdf

Ruch, P., Baud, R., Rassinoux, A., Bouillon, P., & Robert, G. (2000). Medical document anonymization with a semantic lexicon. *Proceedings from AMIA Symposium*, 729–733.

Rudniy, A. (2018). Case study—Data warehouse design for evidence-based research. *IEEE Transactions on Professional Communication*. Forthcoming.

S. 1490 — 111th Congress: Personal Data Privacy and Security Act of 2009. Retrieved from https://www.govtrack.us/congress/bills/111/s1490.

Saeed, M., Lieu, C., Raber, G., & Mark, R.G. (2002). MIMIC II: A massive temporal ICU patient database to support research in intelligent patient monitoring. *Computers in Cardiology, 29*, 641–644. doi: 10.1109/CIC.2002.1166854.

Sang, E.F.T.K., & De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. *Proceedings: Seventh conference on Natural language learning at HLT-NAACL*. Association for Computational Linguistics, 142–147.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T, Ananiadou, S., & Tsujii, J. (2012). BRAT: A Web-based tool for NLP-assisted text annotation. *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 102-107.

Stubbs, A., Kotfila, C., & Uzuner, O. (2015). Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1. *Journal of Biomedical Informatics, 58(S)*, 11–19.

Summary of the HIPAA Security Rule. Retrieved from https://www.hhs.gov/hipaa/for-professionals/security/laws-regulations/index.html.

Sweeney, L. (1996). Replacing personally-identifying information in medical records, the Scrub system. *Proceedings: A conference of the American Medical Informatics Association. AMIA Fall Symposium*, 333–337.

Sweeney, L. (1998). Datafly: A system for providing anonymity in medical data. In T. Lin, & S. Qian (Eds.), *Database security, XI: Status and prospects* (pp. 356–381). Amsterdam: Elsevier Science.

Sweeney, L. (2000). *Uniqueness of simple demographics in the U.S. Population* (Report # LIDAPWP4). Pittsburgh, PA: Carnegie Mellon University.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10*(5), 557–570.

Szarvas, G., Farkas, R., & Busa-Fekete, R. (2007). State-of-the-art anonymization of medical records using an iterative machine learning framework. *Journal of American Medical Informatics Association, 14*(5), 574–580.

Taguchi, K., & Aramaki, E. (2018). Novel location de-identification for machine and human. UISTDA 2018. Retrieved from http://ceur-ws.org/Vol-2068/uistda5.pdf.

Telephone Consumer Protection Act 47 U.S.C. § 227. Retrieved from https://transition.fcc.gov/cgb/policy/TCPA-Rules.pdf.

The Drivers Privacy Protection Act (DPPA) and the privacy of your state motor vehicle record. Retrieved from https://epic.org/privacy/drivers/.

The Privacy Act of 1974 (As Amended). Public Law 93-579, as codified at 5 U.S.C. 552a. Retrieved from http://www.dodig.mil/Portals/48/Documents/Programs/Privacy%20Program/pa1974.pdf?ver=2017-04-14-103528-910.

The Stanford Natural Language Processing Group. Retrieved from https://nlp.stanford.edu/software/.

Troyano, J., Dıaz, V., Enrıquez, F., & Romero, L. (2004). Improving the performance of a named entity extractor by applying a stacking scheme. In C. Lemaitre, C.A. Reyes, & J.A. Gonzalez (Eds.), *IBERAMIA 2004, LNAI 3315*, 295–304.

USA FREEDOM Act of 2015. Uniting and Strengthening America by Fulfilling Rights and Ensuring Effective Discipline Over Monitoring Act of 2015. Retrieved from https://www.congress.gov/bill/114th-congress/house-bill/2048.

USA PATRIOT Act of 2001. Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism Act of 2001. Retrieved from https://www.gpo.gov/fdsys/pkg/BILLS-107hr3162enr/pdf/BILLS-107hr3162enr.pdf.

Uzuner, O., Luo, Y., & Szolovits, P. (2007). Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association, 14*(5), 550–563. Retrieved from https://doi.org/10.1197/jamia.M2444.

Video Privacy Protection Act of 1988. Retrieved from https://epic.org/privacy/vppa/

Wachtler, J., Khalil, M., Taraghi, B., & Ebner, M. (2016). On using learning analytics to track the activity of interactive MOOC videos. In M. Giannakos, D.G. Sampson, L. Kidzinski, & A. Pardo (Eds.), *Proceedings of the LAK 2016 Workshop on Smart Environments and Analytics in Video-Based Learning* (pp.8–17). Edinburgh, Scotland: CEURS-WS. Retrieved from http://ceur-ws.org/Vol-1579/paper3.pdf.

Wellner, B., Huyck, M., Mardis, S., Aberdeen, J., Morgan, A., Peshkin, L., … Hirschman, L. (2007). Rapidly retargetable approaches to de-identification in medical records. *Journal of American Medical Informatics Association, 14*(5), 564–573.

Yoose, B. (2017). Balancing privacy and strategic planning needs: A case study in de-identification of patron data. *Journal of Intellectual Freedom and Privacy, 2*(1) DOI: http://dx.doi.org/10.5860/jifp.v2i1.

Zeide, E. (2016). Student privacy principles for the age of big data: Moving beyond FERPA and FIPPs. *Drexel Law Review*, *8*(2), 339–394.

Zhao, Y., Zhang, K., Ma, H., & Li, K. (2018). Leveraging text skeleton for de-identification of electronic medical records. *BMC Medical Informatics and Decision Making, 18*(1), 1–18.

Zhou, G., Zhang, J., Su, J., Shen, D., & Tan, C. (2005). Recognizing names in biomedical texts: A machine learning approach. *Bioinformatics*, *20*(7), 1178–1190.