AUTOMATED CLINICAL TRANSCRIPTION

FOR BEHAVIORAL HEALTH CLINICIANS

by

Nazmul Hasan Kazi

A thesis submitted in partial fulfillment
of the requirements for the degree

of

Master of Science

in

Computer Science

MONTANA STATE UNIVERSITY
Bozeman, Montana

December 2021

DEDICATION

I am dedicating this thesis work to four persons who mean a lot to me, have helped me in numerous ways to gain the research skills and knowledge, and have made this thesis work possible and successful. First and foremost, to my mother, Mosammat Farida Begum, who has made countless scarifies to educate me. She has always put my education before everything else and has supported me in my every decision. Second, to my father, Kazi Alauddin Ahmed, who has also made countless sacrifices for me and has supported me in my every decision. Without the support of my parents, I could never have even reached the starting point of this research. Third, to Dr. Indika Kahanda who has taught me how to research and helped me greatly to achieve the research skills and knowledge that I have today. Undoubtedly, he has helped me more than anyone else to become a good researcher. Fourth, to Matt Kuntz who has introduced this research concept to us and has helped us greatly to understand the problem, identify and explore our options, plan our objectives and achieve our goals. He has always made himself available whenever we needed help or ran into complications. Both Dr. Indika Kahanda and Matt Kuntz have helped me greatly to stay focused on my path and have contributed greatly to the success of this research.

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to the thesis committee, Dr. Brendan Mumey, Dr. Indika Kahanda, and Dr. Bernadette McCrory, for their time, support, advice, and feedback. I would like to extend my special thanks to Dr. Kahanda for continuously advising and guiding me through my thesis work even after moving to the University of North Florida and to Matt Kuntz (National Alliance on Mental Illness - Montana) for bringing forth this research concept and for his continuous support from understanding the problem to making the project successful. I am also grateful to Cheryl Bristow (National Alliance on Mental Illness - Montana), Dr. Eric Arzubi (Frontier Psychiatry), and Dr. Wade Hill (Headwaters Mental Health) for their support with data. Many thanks to the Gianforte School of Computing (especially Dr. John Paxton, Dr. Mike Wittie, and Dr. Clemente Izurieta), Deborah Chiolero (Office of International Programs), and Donna Negaard (Graduate School) for their support in numerous administrative works. I am also grateful to Mohammad Anani, Nathaniel Lane, Srinivasan Sridhar, Dr. James Becker, Dr. Kenning Arlitsch, Patrick OBrien, and Dr. Jonathan Wheeler (University of New Mexico) for their collaboration on different research projects that provided me with valuable skills and knowledge for this thesis work. I must thank the Pacific Research Platform for its supercomputing nodes which are heavily utilized for data analysis and model training. Special thanks to the US Economic Development Administration and VPREDGE Office for partially funding this project. I am also grateful to my family, friends, and colleagues for their support, company, and feedback.

TABLE OF CONTENTS

TABLE OF CONTENTS – CONTINUED

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABBREVIATIONS

AI	Artificial intelligence
API	Application programming interface
ARRA	American Recovery and Reinvestment Act
AS	Alexander Street
ASEE	American Society for Engineering Education
AUROC	Area under the receiver operating characteristic
BERT	Bidirectional cncoder representations from transformers
CC	Chief complaint
CD	Client details
EHR	Electronic health record
FH	Family history
GPU	Graphics processing unit
HPI	History of present illness
HSU	History of substance use
NAMI Montana	National Alliance on Mental Illness - Montana
NLG	Natural language generation
NLP	Natural language processing
NLTK	Natural Language Toolkit
PDQI-9	Physician Documentation Quality Instrument, 9-item version
PPH	Past psychiatric history
QA	Question answer
RDoC	Research Domain Criteria
ROC	Receiver operating characteristic
RS	Review of systems
SH	Social history
SR	Speech recognition
STDEV	Standard deviation
SVM	Support vector machine

ABSTRACT

Mental health disorder is one of the most common but expensive healthcare conditions in the world. Yet, more than half of all patients go untreated due to various reasons such as lack of access to resources and clinicians. On the other hand, providers rely on Electronic Health Records (EHRs) to compile and share clinical notes, which is a key component of clinical practice, but time-consuming data entry is considered one of the primary downsides of EHRs. Many practitioners are spending more time in EHR documentation than direct patient care, which adds to patient dissatisfaction and clinician burnout.

In this work, we explore the feasibility of developing an end-to-end clinical transcription tool that fully automates the documentation process for behavioral health clinicians. We divide the task into several sub-tasks and primarily focus on the following: 1) extraction and classification of important information from patient-provider conversations, and 2) generation of clinical notes from extracted information. We develop a dataset of 65 transcripts from simulated provider-patient conversations. Then, we fine-tune a transformer language model that shows promising results on personalized data extraction (F1=0.94) and scope for improvement in classification (F1=0.18) of extracted information to EHR categories. Furthermore, we develop a rule-based natural language generation module that formalizes all types of extracted information and synthesizes them into clinical notes. The overall pipeline shows the potential of automatically generating draft clinical notes and reducing the documentation time for behavioral health clinicians by 70-80%. The findings of this work have implications for health behavioral care providers as well as machine learning and natural language processing application developers.

INTRODUCTION

Mental health falls under the umbrella of behavioral health. Mental health condition is one of the five most expensive healthcare conditions in the United States (Soni, 2009). The costs of mental health rose from 35.2 billion in 1996 to 57.5 billion in 2006 and have the highest out-of-pocket payments among all healthcare conditions since 1996. Mental illness has become one of the largest public health burdens in most societies (Diederich and Song, 2014). According to Mental Health America, nearly 50 million (19.86%) American adults experienced a mental illness in 2019, and 27 million (56%) of them did not receive any form of treatment or counseling (Reinert et al., 2020). More than 24% of American adults reported unmet needs in 2019 and the percentage has been increasing since 2011 (Reinert et al., 2020). Over 60% of American youths with major depression did not receive any treatment (Reinert et al., 2020). Mental health conditions and limited access to resources are some of the leading health concerns in the state of Montana, especially across Native American reservations (Kwon and Saadabadi, 2021). In 2019, 89 thousand (51.1%) adults experiencing a mental illness went untreated in Montana and the availability rate of mental health providers was only 1:320 (Reinert et al., 2020).

Health informatics has been accepted as a way to use information technology, communication technology, and computer science to organize and analyze health records to improve healthcare outcomes (Diederich and Song, 2014). Similarly, mental health informatics can be a way to improve mental health by advancing the information management system, increasing mental care in the remote and under-served population, and making mental care more cost-effective (Diederich and Song, 2014). The World Health Organization (WHO) identified the potential of mental health informatics to support the delivery of mental

health care (Rigby et al., 1998). During the course of my thesis work, I primarily focused on exploring the feasibility of automating clinical documentation for behavioral health clinicians.

Clinical documentation is a key component of clinical practice. Clinicians spend significant time and effort to keep detailed records on their patients through clinical documents containing accurate and exact information strictly defining the patient's conditions, diagnosis, and treatment plan (Combs, 2020). Clinicians try to keep the clinical documents clear, consistent, complete, reliable, legible, precise, and timely (Combs, 2020). There are two major reasons behind such descriptive clinical documentation: 1) ensuring consistency of care, and 2) demonstrating clinical care practices for reimbursement and legal protection.

Descriptive clinical documentation ensures consistency of care (Combs, 2020). Clinicians rely on accurate patient information to create and evaluate treatment plans (Combs, 2020). Clinical documents provide the necessary evidence to integrate clinical expertise with patient's unique values and preferences for making medical decisions and monitoring patient health over time. Patients rely on numerous providers and healthcare professionals to stay in good health. Clinicians often need to consider past and existing medical conditions to make the right treatment plan for current problems. Sometimes, patients follow up with a new provider on their past problems. Clinical documentation is commonly used to facilitate inter-provider communication providing the healthcare team necessary information to stay on course and to provide the needed care to the patients. Clinical documents provide quick and easy access to the patient's medical history and current treatment plans with other providers.

Descriptive clinical documentation also demonstrates clinical care practices for reimbursement (Reyes et al., 2017). According to a national health interview survey conducted by National Center for Health Statistics, about 90% of Americans of all ages have some form of insurance (Cohen et al., 2021). Most insured patients need to pay a small deductible and then the insurance providers pay the rest of the bill. To get reimbursed for the service provided to

a patient, healthcare providers need to submit a proof-of-service to the insurance providers. The proof-of-service should contain detailed information about the care administered during a patient's visit. The insurance providers use the provided information to evaluate the claim and to decide the amount of reimbursement. Clinical documents contain the necessary information for the insurance providers, and are used as proof-of-service by the healthcare providers to file medical claims. A well-known phrase, "If you didn't write it, it did not happen.", is often used in medical training to emphasize the necessity of high-quality clinical documentation (Towers, 2013). In reality, poor or incomplete clinical documentation simply results in little to no reimbursement for an already provided service (Towers, 2013). Thus, clinicians put great importance on keeping detailed records of their patients and high-quality clinical documentation (Combs, 2020; Towers, 2013).

Clear clinical documentation also provides protection for the providers against malpractice claims and disciplinary actions (Zierler-Brown et al., 2007). Providers are required to keep complete, timely, and detailed records on their patients and failure to keep or maintain proper patient records is a direct violation against the Code of Federal Regulations, Title 42 § 482.24. The federal government has introduced the False Claims Act (FCA) of 1986 and the Health Care Fraud and Abuse Control program (HCFAC) to combat healthcare fraud and abuse (Rudman et al., 2009). Providers use clinical documentation to prove and justify provided services (Zierler-Brown et al., 2007). If a provided service is not documented properly, at worse, it can be considered as a fraud in the court of law and result in disciplinary actions (Rudman et al., 2009). Moreover, clinical documents are often collected by the hospitals for internal review or by professional associations to ensure quality patient care (Towers, 2013).

Additionally, evidence generated from clinical documentation of various providers also creates a means for healthcare agencies to research large patient populations which contributes to automating medical decision makings and improving the quality of patient

care. Healthcare agencies use this massive database of clinical documentation to expand the available knowledge of evidence-based clinical practice (Casey et al., 2016).

The practice of compiling and sharing descriptive clinical documentation relies upon electronic health record (EHR) technology. An EHR is a digital version of a patient's health record. According to American Hospital Association (AHA), over 96% of hospitals in the United States adopted a certified EHR technology by 2015 (Henry et al., 2016). Clinical records are stored electronically and clinicians use some form of an electronic device to input their clinical notes. EHRs have improved healthcare quality by reducing the number of medical errors related to hand-written illegible medical records (Bates et al., 1999). A study of 1999 showed that adoption of EHR technology reduced medication-related errors by 81% (Bates et al., 1999). The American Recovery and Reinvestment Act (ARRA) of 2009 mandated all public and private healthcare providers and organizations in the United States to adopt and demonstrate "meaningful use" of a certified EHR system by 2015 (Barrett, 2018). Non-compliant providers in the United States received a 1% reduction in their Medicare reimbursements with an additional 1% reduction for every additional year of delinquency, up to 5% in total (Barrett, 2018). The Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009, passed as part of the ARRA, required the Office of the National Coordinator for Health Information Technology to promote the adoption and meaningful use of EHRs which spurred the use of EHR technology in the United States from 9.4% in 2008 to 96% in 2015 (Henry et al., 2016).

Besides reducing medical errors, healthcare providers also benefit from conserving physician time, sharing patient information among healthcare practitioners, and efficiently automating provider workflow by using EHR technologies (King et al., 2014). EHRs are real-time and provide instant and secure access to authorized users. Physicians can access information like lab results as soon as they are inputted/ uploaded in EHR. Depending on EHR technologies, physicians can even access patient information remotely and provide the

necessary support during emergencies. Modern EHR systems alert clinicians of potential medication errors or critical lab values. EHRs are also used to notify the providers/ patients of their future cares, e.g. vaccine appointments or cancer screening. Due to these benefits, EHRs became a cornerstone of improving the healthcare system (Thakkar and Davis, 2006).

EHRs have downsides as well. The first EHR system was developed in the mid-1960s (Atherton, 2011). Since then, many other EHR systems were developed by academic medical centers, the government, and industries (e.g. Epic, Praxis EMR, Cerner, Allscripts, GE Healthcare, etc.). Some EHR systems, e.g. Epic, are smart enough to detect errors related to medication and critical lab results. However, clinicians still use traditional methods, mouse and keyboard, to input their clinical notes in EHRs. In 2015, the American Medical Informatics Association reported time-consuming data entry as one of the major problems in EHRs (Payne et al., 2015). A 2016 study reported that providers spent 27% of their time on direct patient interaction; whereas 49% of their time (almost double) is spent on EHR documentation working 1-2 hours past their normal work hours each night (Sinsky et al., 2016). This shows that poorly designed EHR input methods are dislodging healthcare providers from their primary task - providing healthcare.

The amount of time spent on data entry is not the same across all types of practitioners (Overhage and McCallie Jr, 2020). Behavioral health clinicians, such as psychiatrists, collect most of their patient information over interviews and need to document nearly all information reported during the conversation since this information provides an understanding of patient's symptoms and medical conditions making these interviews very speech intensive (Cruz et al., 2011). Psychiatry residents reported spending approximately 12 hours writing notes and 9 hours reading notes (22 hours in total) per week (Domaney et al., 2018). Using the EHR led to clinician burnout and high emotional exhaustion particularly among psychiatrists (Domaney et al., 2018). Note writing during sessions limits face time and leads to missing important non-verbal cues highly important for correct diagnosis, patient

dissatisfaction, and clinician burnout(Kazi and Kahanda, 2019). If data entry methods can be improved especially for behavioral health clinicians, then the same methods can also be adapted for other practitioners.

Many commercial companies have developed various softwares to improve the data entry process by utilizing cutting-edge speech recognition (SR) technologies. Software like Amazon Transcribe Medical[1], VoiceboxMD[2], iScribeHealth[3], etc. helps psychiatrists to record patient notes through speech dictation. *Nuance Dragon Medical Practice Edition 4* recognizes medical terms from over 60 medical domains, which make note-taking by dictation more streamlined, accurate, and three times faster than typing (Nuance Communications, Inc., 2021). A clinical survey of 2019 reported that 78.8% of the clinicians were satisfied using SR technologies for note-taking and 77.2% of clinicians agreed that SR improves efficiency (Goss et al., 2019). However, SR marginally reduced the documentation time compared to typing (Blackley et al., 2020; Vogel et al., 2015). Documentation time remains one of the major problems against providing health care without a major reduction in data entry time (Payne et al., 2015).

In 2019, DeepScribe Inc.[4] utilized rule-based NLP (Natural Language Processing) and deep learning techniques to develop a proprietary AI (Artificial Intelligence) model that automatically generates clinical notes. It listens to patient-provider conversations and summarizes an entire conversation into a clinical note (Deepscribe Inc, 2021). The model is trainable by a clinician to customize the notes with preferred phrasing and writing preferences. DeepScribe Inc. claims to reduce the time spent per appointment by 80%. However, the model includes nearly everything mentioned during the conversation in the clinical note, which is not ideal for all types of providers. This will likely generate complete

[1] aws.amazon.com/transcribe/medical/
[2] voiceboxmd.com
[3] iscribehealth.com
[4] deepscribe.ai

notes for primary care physicians but long and redundant notes for psychiatric therapists that can be hard and time-consuming to review or research.

Natural Language Generation (NLG) has been widely used by researchers and developers to generate textual documents or reports using machines. NLG is a subfield of natural language processing, artificial intelligence, computational linguistics, and cognitive science (Reiter and Dale, 1997). It is a process of generating *understandable texts* in human languages based on some input data that varies widely from non-linguistic signals to text documents (Reiter and Dale, 1997; Kavlakoglu, 2020). Researchers have used NLG to auto-generate weather forecast messages from graphical weather maps (Goldberg et al., 1994), summaries from statistical data (Iordanskaja et al., 1992), reports from medical information (Buchanan et al., 1995) and physiological signals (Portet et al., 2009), chatbots (Chen et al., 2018; Wolf et al., 2019), chief complaints from discrete variables in EHR (Lee, 2018), and many more. NLG is one of the most important building blocks to automate the process of generating clinical notes.

In this thesis work, we explore the feasibility of completely automating the documentation process for behavioral health clinicians by collecting most of the contents of the clinical notes directly from the audio recordings of patient-provider conversations. In our first study (Kazi and Kahanda, 2019), we investigate the problem of generating a case note given a digital transcript of a conversation. We divide the overall problem into two subtasks: 1) extraction and classification of important information from transcripts, and 2) develop an NLG module for the summarizing of the extracted text. We compile a dataset of 18 transcripts and label them with five EHR categories. We train a Support Vector Machines (SVM) model over this dataset to extract important information from transcripts and it achieves an overall AUROC (area under the receiver operating characteristic) score of 0.76. We also develop a rule-based NLG module with limited capability to formalize the extracted information. **The findings of this study were presented and published**

in *Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019* (Kazi and Kahanda, 2019). We discuss more details of this study in Chapter 2.

Thirteen of the 18 transcripts in the previously mentioned dataset (i.e., 72% of the dataset) are synthetic transcripts. There are a few major drawbacks of using synthetic transcripts to train an information extraction model. As we increase the number of EHR categories and apply the model over transcripts of real patient-provider conversations, it shows poor performance in extracting information. So, in our second study, we generate a new dataset of 65 transcripts that are directly transcribed using a speech recognition tool from conversations that closely resemble real-life patient-provider conversations. We use eight EHR categories to label this new dataset.

We re-train the SVM model (from the first study) on the new dataset and unfortunately, it fails to successfully extract any important information primarily due to the unbalanced nature of the data (i.e. only 8.65% of the samples are labeled with EHR categories while the rest are irrelevant for generating case notes). We also train a Decision Tree model, a Random Forests model, and a Multi-layer Perceptron model but each of those models displays mediocre performance on extracting important information. So, in this study, we divert our attention to transformer-based deep learning to develop a new classification model. We fine-tune a pre-trained BERT (Bidirectional Encoder Representations from Transformers) model using the new dataset. The BERT model performs significantly better than the SVM model and achieves an F1 score of 0.94 on extracting important information and an overall F1 score of 0.18 in classification.

Furthermore, we also continue our work on the NLG module and add a new rule to overcome the limitation of the NLG module from our first study (Kazi and Kahanda, 2019). The improved NLG module can generate "formal" text for all possible scenarios. We discuss the development of the new dataset, fine-tuning of the BERT model, and the improved NLG module in Chapter 3. Overall, we show the potential of our pipeline to automatically generate

clinical notes directly from patient-provider conversations. **We are currently preparing a manuscript describing this work for submitting to *2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).***

In a separate, yet highly related study, we investigated the best practices and improvements for psychiatry transcript annotation. The main goal of the study was to understand and determine the annotation process for mental health transcripts to acquire more reliable results considering human factor elements, which will help psychiatric clinicians and researchers to develop a more accurate AI model for transcript annotation. **The findings of this study were published in *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care (Sridhar et al., 2021)*.** As the second student author of the paper, I was primarily responsible for developing a web-based application that was used to collect the annotations from the participants. The published paper is included in Appendix C.

In 2010, the National Institute of Mental Health (NIMH) launched a new research framework called Research Domain Criteria (RDoC) to investigate mental disorders by integrating research findings across multiple levels of information ranging from genomics and circuits to behavior and self-report (Carcone and Ruocco, 2017). Before RDoC, the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) and the International Classification of Diseases (ICD) are widely used for the classification of biomedical literature (Lilienfeld and Treadway, 2016). After the RDoC initiative, all existing (i.e., annotated with DSM-5 or ICD framework) and future biomedical research articles need to be categorized with RDoC terminologies in addition to DSM-5 and ICD terms to ensure that the researchers are not missing critical insights from numerous articles on the different sides of the categorization divide.

However, manual curation by human annotators can be highly time and resource

consuming. The rate of publication is increasing at an exponential rate (International Society for Biocuration, 2018) and the RDoC framework is comprehensive and complex making the process of curating research articles challenging for brain researchers (Cuthbert and Insel, 2013).

In order to explore solutions to the above challenges, we introduced an *RDoC task*[5] in the BioNLP-OST 2019 workshop[6] for text miners around the world (see Appendix A). The RDoC task consists of two important sub-tasks that are typical for a triage process (International Society for Biocuration, 2018): a) retrieving PubMed abstracts related to RDoC constructs, and b) extracting the most relevant sentence to a given RDoC construct from a known relevant abstract. The RDoC task provided an opportunity for the community to come-together and develop automated and efficient techniques for annotating biomedical literature with RDoC terminology. **We published the findings from this study in *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, 2019 (Anani et al., 2019)***. I was one of two student organizers for the RDoC Task. I was primarily responsible for calculating inter-annotator agreement metrics, establishing baselines, developing a web application to collect as well as validate the participant submissions, evaluating the submissions, comparing the results, and writing the sections of the manuscript related to these tasks. The published paper is included in Appendix A.

In an unrelated study, we investigated the detection of student misconceptions in short answer questions using natural language processing and machine learning. The main goal of the study was to develop an automated software solution that will eliminate the burden of manually analyzing student responses by providing immediate and personalized feedback to the students on their course writing exercises. **The findings of this study were published in *2021 ASEE Virtual Annual Conference Content Access (Becker***

[5]https://sites.google.com/view/rdoc-task/
[6]http://2019.bionlp-ost.org

et al., 2021). I was the only student author of the paper, and I was primarily responsible for fine-tuning a pre-trained transformer language model (BERT-base) to detect sequential misconceptions in responses. The published paper is included in Appendix D.

In another study involving novel natural language processing and machine learning applications, we explored the feasibility of automatically cataloging scholarly articles using semantic categorical keywords. The main objective of this study was to develop an automated system for cataloging the newly added articles with proper subject headings which will not only expedite the availability of new articles but also create an easy path to retrieve the related articles from institutional repositories. **The findings of this study were published in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop (Kazi et al., 2021)***. I was the first author of this paper. The published paper is included in Appendix B.

The rest of this document is organized as follows: the first study on automatically generating clinical documentation using digital transcripts along with the development of the first dataset and the NLG module is described in Chapter 2, and the second study, i.e. the development of the enhanced dataset, use of transformer-based models, and completion of the NLG module is described in Chapter 3. Finally, we summarize our findings and describe potential future works in Chapter 4.

AUTOMATICALLY GENERATING PSYCHIATRIC CASE NOTES FROM DIGITAL TRANSCRIPTS OF DOCTOR-PATIENT CONVERSATIONS

This chapter discusses our first study on generating case notes from a given digital transcript of a doctor-patient conversation. ***This work was published and presented at Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019 (Kazi and Kahanda, 2019)****. What follows is a verbatim copy of the published article. We use the terms Psychiatrists and Behavioural Health clinicians interchangeably.*

Introduction

An electronic health record (EHR) is a digital version of a patient's health record. EHRs were introduced as a means to improve the health care system. EHRs are real-time and store patient's records in one place and can be shared with other clinicians, researchers and authorized personals instantly and securely. The use and implementation of EHRs were spurred by the 2009 US Health Information Technology for Economic and Clinical Health (HITECH) Act and 78% office-based clinicians reported using some form of EHR by 2013 (Hsiao and Hing, 2014).

Presently, all clinicians are required to digitally document their interactions with their patients using EHRs. These digital documents are called case notes. Manually typing case notes is time consuming (Payne et al., 2015) and limits the face-to-face time with their patients, which leads to both patient dis-satisfaction and clinician burnout. Limited face-to-face time is especially disadvantageous for working with mental health patients where the psychiatrist could easily miss a non-verbal cue highly important for the correct diagnosis. Moreover, EHR's usability related problems lead to unstructured and incomplete case notes (Kaufman et al., 2016) which are difficult to search and access.

Due to the above-mentioned downsides of EHRs, there have been recent attempts

for developing novel methods for incorporating various techniques and technologies such as natural language processing (NLP) for improving the EHR documentation process. In 2015, American Medical Informatics Association reported time-consuming data entry is one of the major problems in EHRs and recommended to improve EHRs by allowing multiple modes of data entry such as audio recording and handwritten notes (Payne et al., 2015). Nagy et al. (2008) developed a voice-controlled EHR system for dentists, called *DentVoice*, that enables dentists to control the EHR and take notes over voice and without taking off their gloves while working with their patients. Kaufman et al. (2016) also developed an NLP-enabled dictation-based data entry where clinicians can write case notes over voice and able to reduce the time by more than 60%.

Psychiatrists mostly collect information from their patients through conversations and these conversations are the primary source of their case notes. In a long-term project in collaboration with National Alliance on Mental illness (NAMI) Montana and the Center for Mental Health Research and Recovery (CMHRR) at Montana State University, we envision a pipeline that automatically records a doctor-patient conversation, generates the corresponding digital transcript of the conversation using speech-to-text API and uses natural language processing and machine learning techniques to predict and/ or extract important pieces of information from the text. This relevant text is then converted to a more formal written version of the text and are used for auto-populating the different sections of the EHR form.

In this work, we focus on the back-end of the above mentioned pipeline, i.e. we explore the feasibility of populating sections of EHR form using the information extracted from a digital transcript of a doctor-patient conversation. In order to gather gold-standard data, we develop a human powered digital transcript annotator and acquire annotated versions of digital transcripts of doctor-patient conversations with the help domain experts. As the first step in our two-step approach, we develop a machine learning model that can predict the

semantic topics of segments of conversations. Then we develop natural language processing techniques to generate a formal written text using the corresponding segments. In this paper, we present our preliminary findings from these two tasks; Figure 2.1 depicts the high-level overview of our two-step approach.

Previous studies most related to our work are (1) Lacson et al. (2006) predicting semantic topics for medical dialogue turns in the home hemodialysis, and (2) Wallace et al. (2014) automatically annotating topics in transcripts of patient-provider interactions regarding antiretroviral adherence. While both studies successfully use machine learning for predicting semantic topics (albeit different topics to ours) they do not focus on the development of NLP models for text summarization (i.e. formal text generation).

The rest of the paper is structured as follows. We describe our two-step approach, data collection and processing, machine learning models and natural language processing methods in chapter 2. In chapter 3, we report and discuss the performance of our methods. We summarize our findings, discuss limitations and potential future work in chapter 4.

Methods

Approach

As depicted in Figure 2.1, we divide the task of generating case notes from digital transcripts of doctor-patient conversations into two sub tasks: (1) using supervised learning models to predict semantic topics for segments of the transcripts and then (2) using natural language processing models to generate a more formal (i.e. written) version of the text which goes in to the corresponding section of the EHR form.

These semantic topics are suggested by the domain experts from NAMI Montana and correspond to the main sections of a typical EHR form. They are (1) Client details: personal information of a patient, such as name, age, birth date etc., (2) Chief complaint: refers to the information regarding a patient's primary problem for which the patient is seeking medical

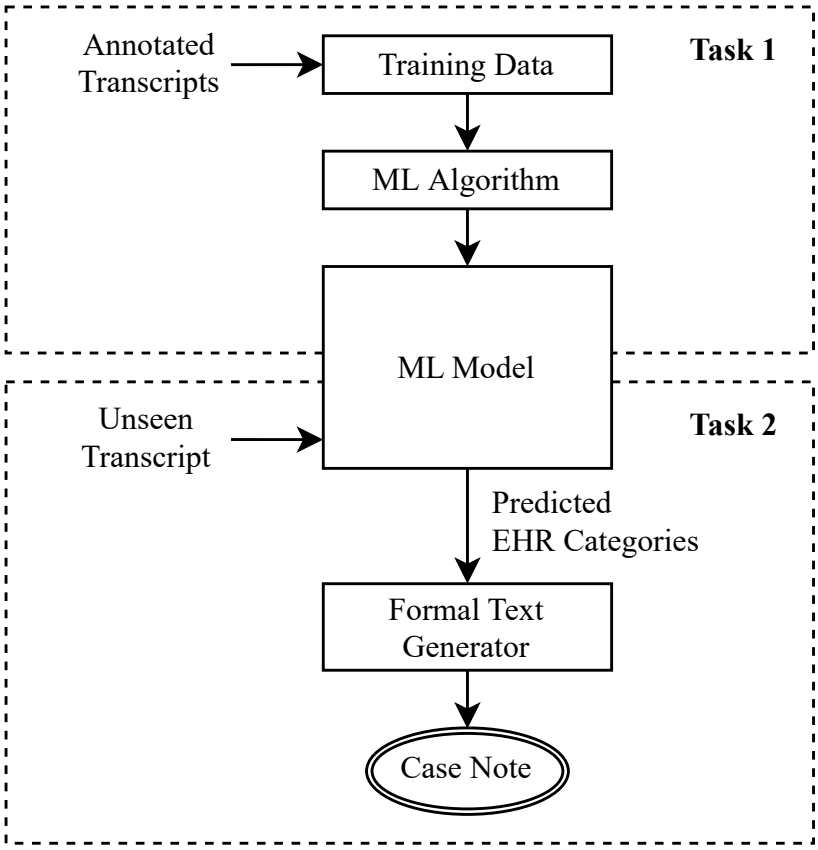

Figure 2.1: High-level overview of our approach. Task 1: Predicting EHR categories. Task 2: Formal text generation. ML: Machine Learning. EHR: Electronic Health Record.

attention., (3) Medical history: any past medical condition(s), treatment(s) and record(s), (4) Family history: indicates medical history of a family member of the patient, and (5) Social history: refers to information about patient's social interactions, e.g. friends, work, family dinner etc. We call these semantic categories "EHR categories" interchangeably. The *formal text* is essentially the summary text that the clinician would write or type into the EHR form based on the interaction with the patient.

Transcripts of Doctor-Patient Dialogue

Our raw dataset is composed of 18 digital transcripts of doctor-patient conversations and covers 11 presenting conditions. The presenting conditions are Attention-deficit/

hyperactivity disorder (ADHD), Alzheimer's disease, Anger, Anorexia, Anxiety, Bipolar, Borderline Personality Disorder (BPD), Depression, Obsessive Compulsive Disorder (OCD), Post Traumatic Stress Disorder (PTSD) and Schizophrenia. All transcripts are labeled with speaker tags "Doctor:" and "Patient:" to indicate the words uttered by each individual.

Thirteen of these transcripts are *synthetic* in that they are handwritten (i.e. typed) by a domain expert from NAMI Montana who has years of experience working with mental illness patients. Hence, each synthetic transcript represents a real case scenario of conversation between a patient (suffering from one of the presenting conditions mentioned above) and a psychiatric doctor/ clinician who verbally interviews the patient in a 2-person dialogue set up. Table 2.1 reports summary statistics.

	Synthetic Transcripts			AS Transcripts		
Property	Total	Mean	STDEV	Total	Mean	STDEV
No. Sentences	1930	148.4	55.6	1390	278.0	74.9
No. Questions	513	39.4	19.8	188	37.6	7.1
No. Dialogue turns	861	66.2	40.0	684	136.8	55.0
No. Sentences spoken by the Doctor	751	57.7	30.0	581	116.2	44.3
No. Sentences spoken by the Patient	1179	90.6	33.2	809	161.8	60.5

Table 2.1: Summary statistics on 13 synthetic transcripts vs. 5 AS (Alexander Street) transcripts.

Rest of the five transcripts are part of *Counseling & Therapy* database[1] from the Alexander Street website. Hence, we refer to them as AS transcripts for the rest of the paper. Each of these AS transcripts is generated from a real-life conversation between a patient and a clinician. Majority of these transcripts cover multiple mental conditions.

In order to annotate transcripts using semantic topics mentioned above, we develop a human-powered transcript annotator as shown in Figure 2.2, a responsive web application, that takes digital transcripts as input, breaks down each transcript into segments where each

[1]https://search.alexanderstreet.com/health-sciences/counseling-therapy

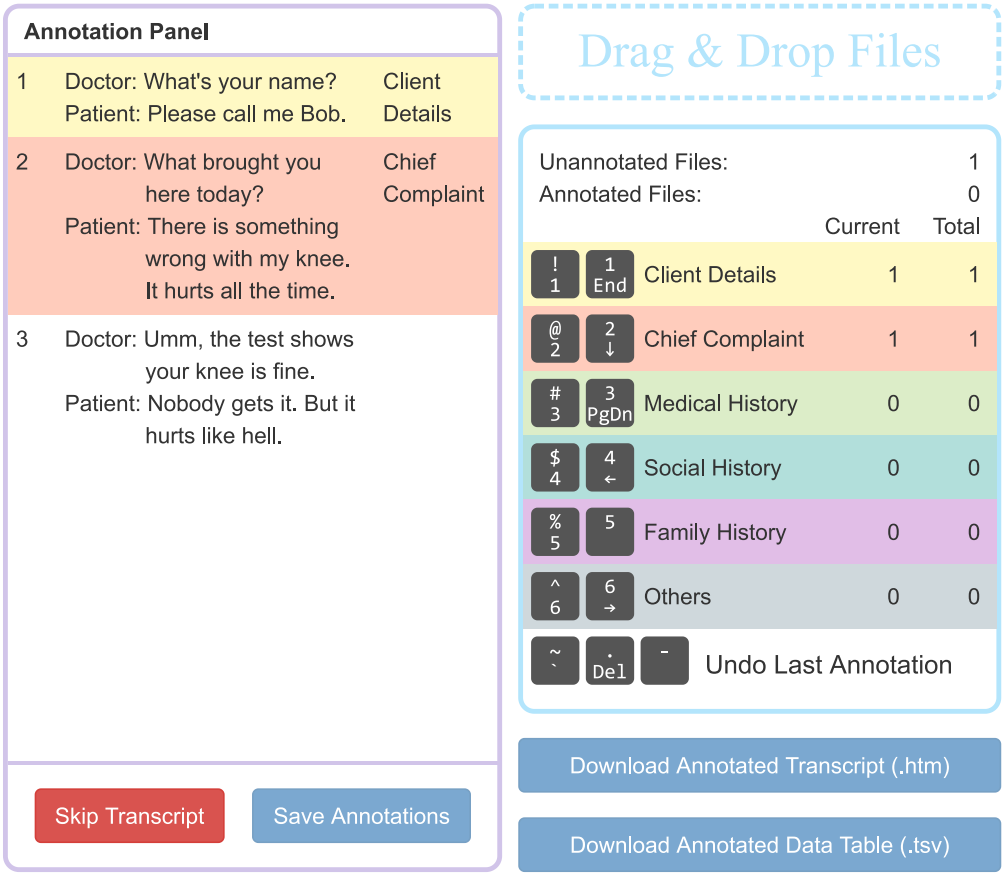

Figure 2.2: Screen shot of the human-powered transcript annotator. Left panel displays an example transcript while the semantic concepts are shown on the right.

segment starts with a speaker tag (Doctor: or Patient:) and generates samples by pairing each doctor segment with the followed by patient segment. The application displays the generated samples, from one transcript at a time, in the same order as they appear in the transcript and allows the user to annotate them with one of the six semantic topics.

A group of three annotators including two domain-experts from NAMI Montana use the above annotator tool to single-annotate (through collaboration) all 18 transcripts. As highlighted in Figure 2.2, annotations are added at the *conversation pair* level. We define the conversation pair as the entire text associated with a consecutive pair of "Doctor:" and "Patient:" speaker tags. Each conversation pair is annotated with one of the five topics

(i.e. EHR categories). These labels are based on the main focus/ subject/ topic of the corresponding conversation pair as judged by the expert annotators. Any conversation pair that was found to be irrelevant to the five categories is annotated with a new category called "Others". Conversation pair level annotations eliminated the challenges in annotating a question or an answer on their own without the proper context provided by the preceding/ following sentences.

Task 1: Predicting EHR Categories

In this task, we use the annotated digital transcripts to generate the training data to train supervised classification models using two different approaches. These two approaches mainly differ in how the transcripts were segmented into examples (i.e. training instances) for generating the training datasets as described in the sections 2 and 2. Regardless of the approach, we label the examples with one of the six class labels analogs to the semantic topics (EHR categories): Client Details, Chief Complaint, Family History, Social History, Medical History and Others.

Training Data - Model 1 In this approach, we build a training dataset by taking a conversation pair as a single example (i.e. instance). Each example contains at least two sentences where the first sentence is spoken by the doctor and the second sentence is spoken by the patient. The class label for each example is the corresponding annotation from the original transcript; this results in six class labels. A short examples of the training dataset and distribution of class labels are reported in Tables 2.2 and 2.3.

Segmenting the transcripts into training examples in this fashion is convenient because there is a one-to-one mapping between the semantic topics in the original annotated transcripts and the class labels of the examples; additional reconciliation is not needed. However, sometimes, the doctor or the patient talks about more than one topic (inside the same conversation pair). For example, although example 2 in Table 2.2 is labeled with

No.	Example	Class Label
1	Doctor: How many voices do you hear? Patient: Two. They talk all the time.	Chief Complaint
2	Doctor: Your record shows that you take antidepressants pills regularly. Do you hang out with your parents, co-workers or friends? Do you talk to them? Patient: Sometimes I hang out with my mom. Yes, I talk to my co-workers but only for work. I used to have a friend who moved couple months ago and we don't talk anymore.	Social History

Table 2.2: Examples in Model 1 training data.

Class Label	Synthetic		All
	Model 1	Model 2	
Chief Complaint	309	870	1746
Client Details	32	88	198
Family History	28	101	149
Medical History	34	74	85
Others	12	174	264
Social History	19	51	110
Total	434	1358	2552

Table 2.3: Distribution of class labels in training data. All: represents Model 2 training data with all 18 transcripts.

Social History, the conversation pair is composed of information relevant to both the medical history and social history. Therefore, segmenting the transcript to smaller pieces could be more beneficial for improved overall performance. This is the motivation for the second approach mentioned in the next section.

Training Data - Model 2 In this approach, we use a finer-level granularity (than conversation pairs) for segmenting the transcripts for generating training examples. We start with the Model 1 training data and tokenize the text of each example at the sentence level

by identifying the sentence boundaries using sentence tokenizer in NLTK[2]. We first assign labels to each sentence based on the class label of the original source (i.e. conversation pair). Then, one of the human annotators manually reviewed the class labels and makes corrections if needed.

However, labeling at the sentence-level is also challenging because the information that defines the topic (class label) lies in the question and is sometimes followed by a short answer, e.g. Table 2.4 example 1. We also observe the opposite scenario where the answer holds the context, e.g. Table 2.4 example 2, and scenarios where the information lies in both the question and the answer, e.g. Table 2.4 example 3. So, it is understood that without pairing the questions with their corresponding answers (or being aware of the context provided by the question or the answer), it is challenging even for human annotators to label these sentences individually. However, We also observe that a question is commonly followed by its corresponding answer in the form of a non-interrogative sentence. Therefore, we use the following approach to overcome the above challenge.

We first combine the grammatical rules of the English language in forming a question (British Council, 2019) and spaCy[3], an industrial-strength natural language processing API, to identify the questions in the transcript. Then, to preserve the context, we pair

[2]https://www.nltk.org/
[3]https://spacy.io/

#	Question-answer pair	Class Label
1	Doctor: How old are you? Patient: 23.	Client Details
2	Doctor: Who do you take? Patient: I take Ibuprofen.	Medical History
3	Doctor: What is your name? Patient: Name is a game.	Chief Compliant

Table 2.4: Question-answer pair dependency.

No.	Example	Class Label
1	How many voices do you hear? Two.	Chief Complaint
2	They talk all the time.	Chief Complaint
3	Your record shows that you take antidepressants pills regularly.	Medical History
4	Do you hang out with your parents, co-workers or friends? Do you talk to them? Sometimes I hang out with my mom.	Social History
5	Yes, I talk to my co-workers but only for work.	Social History
6	I used to have a friend who moved couple months ago and we don't talk anymore.	Social History

Table 2.5: Examples in Model 2 training data.

each question with the following non-interrogative sentence and combine them into a single example. In other words, Model 2 training instances can be single sentences or a conversation pair or anything in between. Several examples of Model 2 dataset is shown in Table 2.5. These examples correspond to the Model 1 examples depicted in Table 2.2.

Machine Learning Models To explore the feasibility of classifying information from digital transcripts, we train separate supervised learning classifiers using both training datasets (i.e. Model 1 and Model 2). Specifically, since each instance is annotated with exactly one class label (out of six), we model this as a multi-class problem and use the one-vs-rest (Bishop, 2006) classification strategy.

We apply Support Vector Machines (SVMs) as our machine learning algorithm (which was found to be the best performer in an initial study in comparison with a few other popular machine learning algorithms: k-Nearest Neighbors, Naïve Bayes, Decision Tree, Neural Networks – data not shown). We use stop word removal and lemmatization for pre-processing and Bag-of-Words model for feature extraction. We use sci-kit learn (Pedregosa et al., 2011) python machine learning library for implementing these models. For our preliminary experiments reported in this paper, we do not use any model checking or parameter tuning and use default settings.

Task 2: Formal Text Generation Due to the error-prone nature of Model 1 training data described above, we exclusively use Model 2 training data for the formal text generation. The high-level idea is that in order to generate a case note for an unseen transcript, we first segment the transcript at the Model 2 granularity and predict the EHR categories using the Model 2 classifier. Then instances are grouped based on their predicted EHR categories. Generating case notes with sentences as they appear in the transcripts (i.e. verbatim) will result in redundant case notes that will be difficult to search for important information. An assertive sentence generated by gathering information from a question-answer pair will be easier to read and concise. Therefore, for each category, a formal written version of the text is generated using the method described below. We ignore 'Others' category in our current setup because they represent irrelevant information and any information under this class is likely not important for case note.

In order to generate formal text from an instance, the entire text needs to be rewritten using an assertive sentence, subject in third person singular form, correct tense, verb form and sentence structure. We concatenate each piece of formal text within the category to form a paragraph. Thus, our method results in generating a case note composed of five paragraphs corresponding to the first five EHR categories.

As illustrated in Figure 2.3, our method generates formal text in several steps. As mentioned above, a sample can be either a sentence or a question-answer pair (as depicted in Table 2.5). First, we identify the number of sentences in the example text. Examples composed of a single sentence (e.g. Table 2.7, examples 1-3) requires minimal processing to generate formal text. We use part-of-speech tagging from python module spaCy to identify the subject, main verb and the auxiliary verb(s) of the sentences. If the subject is a first (I) or second person (you), the subject is replaced with the third person singular form (he/she). Clinicians typically collect personal information, such as name, gender and contact information, prior to their conversation or appointment and so they can be fed into our model

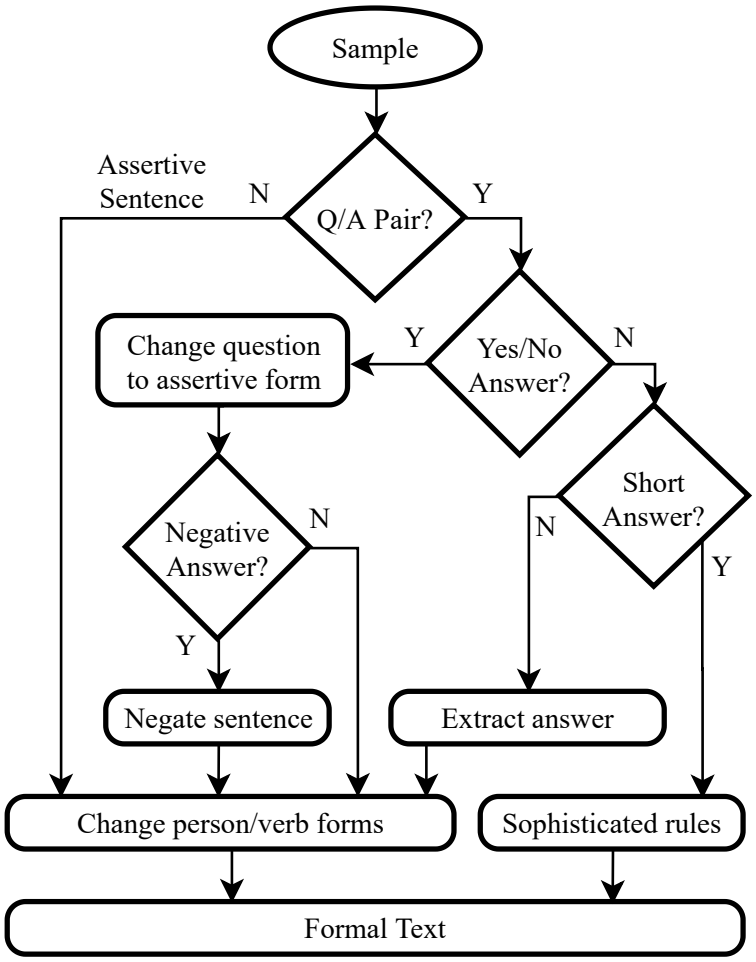

Figure 2.3: An overview of formal text generation steps.

as input to generate accurate case notes.

If the sentence contains auxiliary verb(s), the first auxiliary verb is replaced with its third person singular form, e.g. *am* with *is*, and the second auxiliary verb, if any, and the main verb are kept as they are. If the sentence does not contain any auxiliary verbs, the proper form of the main verb depends on the tense of the sentence. If the sentence is in the present tense, the main verb is replaced with its third person singular form, e.g. *run* with *runs*. For sentences in the past tense, the main verb is kept unchanged since the form of the verb is the same for all persons, e.g. *took*. A sentence in future tense contains at least one

auxiliary verb, *shall* or *will*, and therefore our method processes the sentence as a sentence in the present tense; there is no need to add any additional functionality to cover this tense.

If an instance is composed of multiple sentences, the last sentence is always a non-interrogative sentence and is the answer to the question posed in the very first sentence. In this case, the formal text depends on both the question and the answer. If the answer starts with an affirmation or negation word (e.g. yes, no, yeah, never), the question is changed to an affirmative or negative sentence, respectively, and the assertive sentence is added as a separate sentence after removing the leading affirmation or negation word (e.g. Table 2.7, examples 4-5). If the answer does not start with any affirmation or negation word, the answer is further analyzed to see whether it is a short answer. If not, the question text is ignored and the answer text is returned as the formal text (e.g. Table 2.7, example 6).

In the case of short answers, an answer alone does not provide the full context to construct the formal text and we need to rely on both the question and the answer. For e.g. the wh- questions (e.g. when, who) are usually followed by a relatively short answer that requires context from the question text as well. This required more sophisticated rules and we are presently working on generating formal text for this scenario. Examples and the intended "ideal" formal text for them are given in Table 2.8.

While generating formal text, all first and second person pronouns, regardless their position, are replaced with their third person singular form and the verbs are also replaced with its third person singular form, where applicable. Regular expressions are used to remove leading words (e.g. ok, right, yes, and, but, hmm) from the assertive sentences that have no importance to be included in the formal texts. This functionality was implemented using NodeBox[4] Python library.

[4]https://www.nodebox.net/code/index.php/Linguistics

Experimental Setup and Metrics

In terms of Task 1, we evaluate our supervised machine learning models using 5 fold stratified cross-validation and the performance is reported using the AUROC (Area Under the ROC Curve) scale (Bewick et al., 2004). A score of 1 corresponds to the performance of an ideal classifier whereas a score of 0.5 relates to the performance of a random classifier. Because Task 2 (formal text generation aspect) of the project is a work-in-progress, we highlight the scenarios that our model is able to handle and mention the more challenging scenarios in future work.

Results and Discussion

In an initial experiment, we assessed the performance of Model 1 and Model 2 training data using the 13 synthetic transcripts. According to our preliminary results, SVMs with linear kernel performs the best with a macro-average AUROC score of 79% for Model 1. For Model 2, the SVMs classifier achieves a macro-average AUROC score of 81%. However, note that these numbers are not directly comparable because Model 1 training instances are different from that of Model 2. Still, this suggested that Model 2 is superior in performance. This is intuitive because Model 2 training data is a more refined dataset as described

Class Label	AUROC	STDEV
Chief Complaint	0.74	0.02
Client Details	0.73	0.04
Family History	0.77	0.04
Medical History	0.78	0.07
Others	0.84	0.03
Social History	0.67	0.06
Macro-average	0.76	

Table 2.6: Performance of Model 2 training data using all transcripts (13 artificial and 5 AS). Performance collected through 5-fold cross validation, repeated 10 times.

previously. This observation, coupled with the fact that Model 2 data are more conducive to formal text generation, we used Model 2 training data for the rest of the experiments.

Next, we assessed the performance of Model 2 using all the transcripts (i.e. 13 synthetic

No.	Example	Generated Formal Text
1	I do not seem to be coping with things.	He does not seem to be coping with things.
2	I woke up about 4 am last night.	He woke up about 4 am last night.
3	My sister said I should come.	His sister said he should come.
4	Do you have any sort of hallucination and delusion? No.	He does not have any sort of hallucination and delusion.
5	Has this been going on for some time? Yeah, a few months really.	This has been going on for some time. A few months really.
6	Ok, so what is brought you here today? My sister's noticed, I am just a bit fed up really with some mood swings.	His sister's noticed, he is just a bit fed up really with some mood swings.

Table 2.7: Formal Text Generation: example inputs and the generated text.

No.	Example	Ideal Formal Text
1	Where do you work? A shop near the mall.	He works in a shop near the mall.
2	When did you wake up last night? It was before 4.	He woke up before 4 last night.
3	When did that happen? Then I was 10.	That happened when he was 10.
4	How often do you exercise? Not that much, I play basketball on Mondays and go for a run on Wednesdays and Saturdays.	He does not exercises much. He plays basketball on Mondays and goes for a run on Wednesdays and Saturdays.
5	Which color shall we use? Red, use red.	We shall use red.
6	In what way does he push her? Not like with hands, just ignores her to make her mad.	He does not push her with hands, just ignores her to make her mad.

Table 2.8: Formal Text Generation: challenging examples (requiring sophisticated rules) and their *ideal* formal text.

and 5 AS transcripts). There is a clear performance dip (0.81 vs. 0.76) when the AS transcripts are added to the training data. This is intuitive because we believe the AS transcripts may have lead to data that is harder to generalize for the classifiers. The reason is that the majority of them is associated with multiple presenting conditions and hence the content of the questions and answers may be broader than synthetic transcripts. Also, the language characteristics between the synthetic and AS transcripts have a noticeable difference according to Table 2.1. However, this provides valuable insight into the importance of the robustness of the classifier. In other words, caution must be exercised when synthetic data are used for training machine learning models. Note that we did not conduct a separate experiment with only the AS transcripts because the number of examples for some of the ill-represented classes were deemed inadequate.

We observe that the performance for the individual semantic topics (EHR categories) fall in the range of 0.67 (Social History) and 0.84 (Others) as depicted in Table 2.6. But there is no correlation between the class distribution and the performance as evident from Table 2.3. Overall, these numbers suggest that the words of the transcript are reasonably informative for differentiating EHR categories but there is definitely room for improvement. One such improvement may come from focusing on the *type* of the words in addition to their lexical value. This view is supported by the top 5 tokens identified by the classifier

Class Label	Top five features
Chief Complaint	percent, stuff, feeling, number, feel
Client Details	meet, learned, write, pack, style
Family History	cousin, supportive, dad, married, family
Medical History	teen, asthma, dr., prozac, advair
Others	lab, ok, let, right, thank
Social History	comment, wellbutrin, racist, share, friend

Table 2.9: List of top five features per category used by the machine learning classifier.

as the most important tokens for each category (Table 2.9). For example, many of the top words for Family History are names of family members. We also emphasize that the performance reported is from models that work with BoW features and default parameter values, suggesting that the use of a comprehensive feature/ model selection procedure would likely yield better results.

As mentioned above, our formal text generation module is able to handle the scenarios listed in Table 2.7. However, instances in which the context lies in both the question and the answer (e.g. Table 2.4 example 3) are clearly more challenging and hence would require sophisticated rules. In such cases, the challenge is to extract information from both the question as well as the answer and to form an assertive sentence using the combined information. We are currently working on this scenario. Table 2.8 depicts examples from this scenario and the ideal formal text that must be generated.

Conclusion and Future Work

In this work, we focus on the problem of automatically generating case notes from digital transcripts of doctor-patient conversations, using a two-step approach: (1) predicting EHR categories and (2) generating formal text. On the task of predicting semantic topics for segments of the transcripts, we develop a supervised learning model while for the subsequent task of generating a formal version of the text from those segments, we develop a natural language processing model. According to preliminary experimental results obtained using a set of annotated synthetic and real-life transcripts, we demonstrate that our two-step approach is a viable option for automatically generating case notes from digital transcripts of doctor-patient conversations.

However, as noted previously, this is an ongoing project. The immediate attention is paid to handling the case of generating case notes for examples related to short answers given in Table 2.8. Due to the complexity of this scenario, sophisticated rules that make use

for entities identified in the text must be utilized. We plan to transcribe authentic doctor-patient interactions and train a new classification model using these transcripts. We also intend to build a prototype and send it to clinicians for testing using PDQI-9 (Stetson et al., 2012) to check the quality of our generated case notes.

CURAVOICE: AN END-TO-END AUTOMATED CLINICAL TRANSCRIPTION SYSTEM FOR BEHAVIOURAL HEALTH CLINICIANS

This chapter discusses our latest approach of generating clinical notes directly from patient-provider conversations. We are currently preparing a manuscript to submit this work to the 2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics.

Introduction

Clinical documentation is a key component of clinical practice. Clinicians put their full effort to keep detailed records on their patients for evidence-based patient care and to get reimbursed for the care provided. Clinical documents are used to facilitate intra and inter-provider communication, provide the healthcare team necessary information to stay on course, create/ evaluate treatment plans, query patient history, research a large patient population, and for numerous other purposes. According to American Hospital Association (AHA), over 96% of hospitals in the United States adopted a certified Electronic Health Record (EHR) technology by 2015 (Henry et al., 2016). EHRs were introduced to improve healthcare quality by reducing the number of medical errors related to hand-written illegible medical records (Bates et al., 1999). A 1999 study showed that the adoption of EHR technology reduced medication-related errors by 81% (Bates et al., 1999). The American Recovery and Reinvestment Act (ARRA) of 2009 mandated all public and private healthcare providers and organizations to adopt and demonstrate "meaningful use" of a certified EHR system by 2015 (Barrett, 2018). Non-compliant providers in the United States received a 1% reduction in their Medicare reimbursements with an additional 1% reduction for every additional year of delinquency, up to 5% in total (Barrett, 2018). The Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009, passed as part of the

ARRA, required the Office of the National Coordinator for Health Information Technology to promote the adoption and meaningful use of EHRs which spurred the use of EHR technology in the United States from 9.4% in 2008 to 96% in 2015 (Henry et al., 2016).

Clinicians use traditional methods, mouse and keyboard, to input their clinical notes in EHRs. In 2015, American Medical Informatics Association reported time-consuming data entry as one of the major problems in EHRs (Payne et al., 2015). A 2016 study reported that providers spent about 1.8 times more time in EHR documentation than on direct patient care (Sinsky et al., 2016). This clearly indicates that the EHR input methods need to be improved to keep the healthcare providers engaged in their primary task - providing care.

The data entry time varies among different types of clinicians (Overhage and McCallie Jr, 2020). Behavioral health clinicians, e.g. psychiatrists, collect most of their patient information over interviews and information reported during the conversation provides an understanding of patient's symptoms and medical conditions making the interviews very speech intensive (Cruz et al., 2011). Psychiatry residents reported spending 22 hours per week reading and writing clinical notes (Domaney et al., 2018) which is much higher than the average documentation time reported by other clinicians (Overhage and McCallie Jr, 2020). Note writing during sessions limits face time and leads to missing important non-verbal cues highly important for correct diagnosis, patient dissatisfaction, and clinician burnout. In this work, we aim to improve EHR data entry methods for behavioral health clinicians, which can be adapted for other practitioners as well.

Many commercial companies have utilized cutting-edge SR technologies to improve the data entry process. Software like Amazon Transcribe Medical[1], VoiceboxMD[2], iScribeHealth[3], etc. helps clinicians to input their clinical notes through speech dictation. *Nuance Dragon Medical Practice Edition 4* recognizes medical terms from more than 60

[1]aws.amazon.com/transcribe/medical/
[2]voiceboxmd.com
[3]iscribehealth.com

specialized medical vocabularies to make dictation for clinicians more streamlined, accurate and notes taking three times faster than typing (Nuance Communications, Inc., 2021). A clinical survey of 2019 reported that 78.8% of the clinicians were satisfied using SR technologies for note-taking and 77.2% of clinicians agreed that SR improves efficiency (Goss et al., 2019). However, SR marginally reduced the documentation time compared to typing (Blackley et al., 2020; Vogel et al., 2015). Without a major reduction in data entry time, documentation time remains one of the major problems against providing health care (Payne et al., 2015).

In 2019, DeepScribe Inc.[4] launched a commercial product called DeepScribe that automatically generates clinical notes by listening to patient-provider conversations. It uses deep learning and rule-based NLP (Natural Language Processing) to summarize an entire conversation into a clinical note which reduces the time spent per appointment by 80% (Deepscribe Inc, 2021). Clinicians can also train the model to customize their notes. However, the model includes nearly everything reported during the conversation in the clinical note which generates complete notes for primary care physicians but likely long and redundant notes for clinicians like psychiatric therapists making the notes hard and time-consuming to review and research.

In this project, we aim to completely automate the documentation process by collecting most of the information for the clinical notes from the patient-provider conversation. Instead of typing or dictating a note from scratch, a clinician starts with a draft clinical note that is auto-filled with mentioned information during the conversation. Then, the clinician can add, edit or remove information from the draft note as necessary. Our model does not draft clinical notes in the same fashion for all clinicians because each clinician has his/her own preferences for categorizing information. Our model learns the categorization from a clinician's daily practices and drafts the notes accordingly. The model also investigates the

[4]deepscribe.ai

changes made to a clinical note to stay adaptive to the clinician's writing style. Once the model is well-trained (i.e., the learning rate reaches close to zero), we estimate a 70%-80% reduction in documentation time.

Methods

Approach

The overall pipeline of generating clinical documentation directly from patient-provider conversations is outlined in Figure 3.1. A patient-provider conversation is recorded using microphone(s). An existing speech-to-text API is used to transcribe the audio recording into a digital transcript. Then, the classification model extracts important information from the transcript and labels them with EHR categories. The NLG module formalizes the extracted information, group them into paragraphs based on their labels (i.e., EHR categories) and

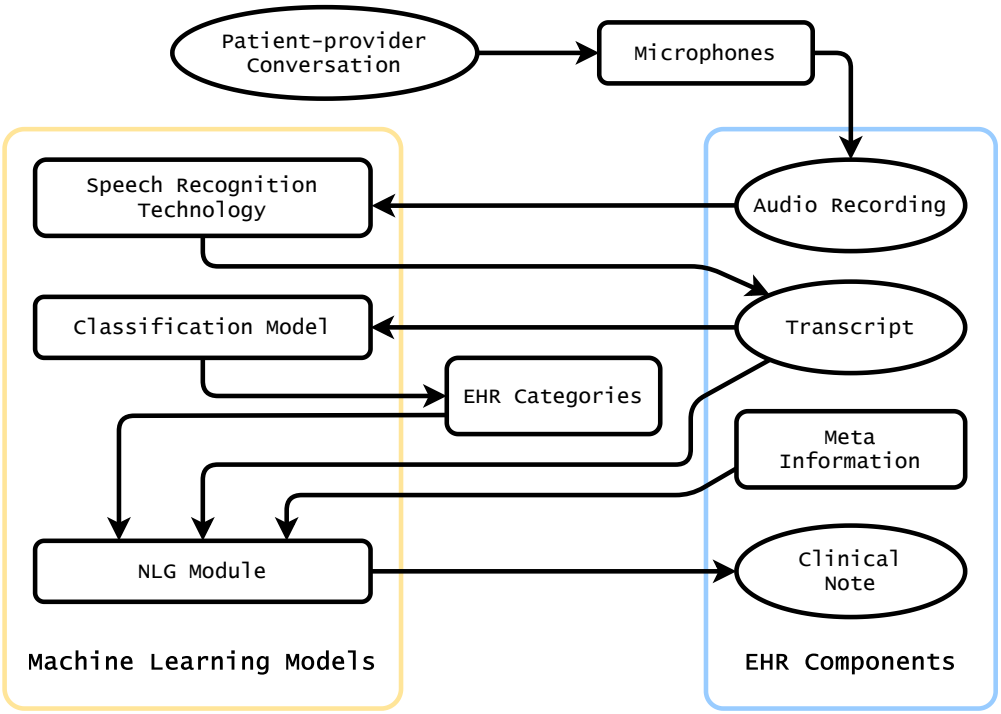

Figure 3.1: A high-level overview of our approach.

return it as a clinical note. Along with the clinical note, a provider can choose to store the audio recording and the transcript for future use or reference.

Audio Recordings and Transcripts

In our preliminary work, we use a dataset composed of thirteen synthetic transcripts (i.e., hand-written by a domain expert) and five natural transcripts hand-transcribed from the real-life patient-provider conversation by professional human transcriber(s). There are a few major drawbacks of using synthetic transcripts for this research. A synthetic transcript written by a domain expert does not always truly imitate a real patient-provider conversation. Often, sentence formations are quite different than the sentences spoken during a natural conversation. A model trained over synthetic transcripts shows poor performance in classifying sentences of natural transcripts (data not shown). This raises the necessity of developing a new dataset with natural transcripts.

Comprising a dataset from real patient conversations for research has some downsides. Conversations between a patient and a behavioral health clinician are protected under the Health Insurance Portability and Accountability Act of 1996 (HIPAA) that protects the privacy and security of certain patient information. Storage, access, sharing, analysis, and use of such protected data must be HIPAA compliant as well. Any violations of HIPAA can result in criminal penalties by the Department of Justice (Vanderpool, 2012). To keep the flexibility of experimenting with and use of components, e.g. transcription APIs, and resources, e.g. GPU accelerated compute nodes, that are not HIPAA compliant but have the potential of providing promising results or accelerating the research, we put our best effort to generate a dataset that closely imitates real patient conversations.

The multimedia resource of Alexander Street Press on behavioral and mental health programs[5] includes more than 6,000 transcripts from real-life therapy sessions covering over

[5] alexanderstreet.com/discipline/behavioral-mental-health

147 presenting conditions to provide unprecedented access to researchers. We use these transcripts as the foundation of our dataset. These transcripts are contributed by 264 publishers and are not transcribed in a common format. Most of these transcripts are transcribed by human transcribers and include non-verbal information, e.g. impression and action of patients, that does not get recognized by speech recognition APIs. Since these transcripts do not truly represent the transcripts generated by a speech recognition API, we can not use them as-is for this research. Instead, we employ six senior psychology students from Montana State University to enact the conversation captured by these transcripts. Then, we transcribe the audio recordings using a speech recognition (or speech-to-text) API.

We do not use any specific criteria for selecting the transcripts for enactment. We choose them at random with a mindset of covering a variety of mental conditions and different session lengths. The students work in pairs, one enacting the patient and another enacting the provider, and generate audio recordings of about 30 hours from 65 transcripts. We record each speaker in a separate (audio) channel to eliminate the need of using AI-based solutions for speaker diarization. We use the standard model of Google Cloud Speech-to-Text API [6] with separate recognition per channel to transcribe the audio recordings.

EHR Categories

We adapt the EHR categories from the *American Psychiatric Association Practice Guidelines for the Psychiatric Evaluation of Adults*. We use the following EHR categories from the practice guidelines: Client Details (CD), Chief Complaint (CC), History of Present Illness (HPI), Past Psychiatric History (PPH), History of Substance Use (HSU), Social History (SH), Family History (FH), and Review of Systems (RS). Clinicians put only the important information mentioned during a session in the clinical note under one of the EHR categories. As we can see in Table 3.1 that less than 9% of the contents from the transcripts

[6] cloud.google.com/speech-to-text

are included in the clinical notes. We add a new category called *Others* to represent these contents that are irrelevant and held out of the clinical notes.

Training Data

We develop an application to generate the training data from the transcripts. A screenshot of this application is shown in Figure 3.2. The application shows a transcript on the left side and an empty EHR form (i.e., clinical note) on the right side of the screen. The EHR form includes a content box for each EHR category. A user can drag and drop

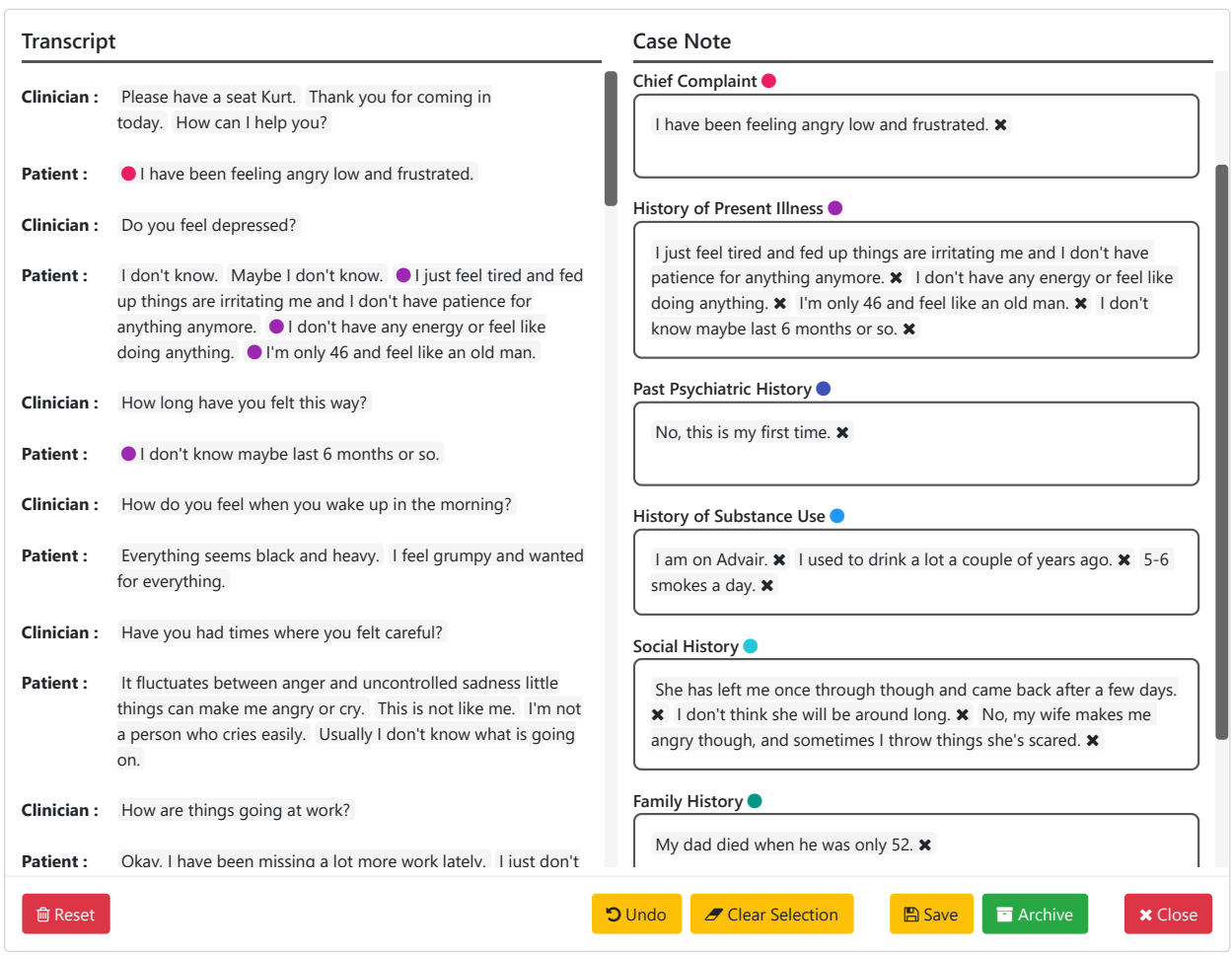

Figure 3.2: The web interface of the application is used by the students to annotate the transcripts.

any number of sentences from the transcript to any of the content boxes. When the user drops a sentence into a content box, the sentence is labeled with the corresponding EHR category and added to the database. This way, a user can easily annotate the important information of a transcript through the process of populating the clinical note by dragging and dropping the sentences to the correct EHR category. We ask the same psychology students to generate clinical notes for all 65 transcripts. Each student has prior experience in writing clinical notes. Therefore, we let the students populate the clinical notes based on their understanding of the EHR categories without any training or instruction. The count and distribution of contents from the transcripts across all categories are listed in Table 3.1.

Category	Count	Percentage
CD	332	1.58%
CC	564	2.96%
HPI	451	2.15%
PPH	121	0.58%
HSU	18	0.09%
SH	145	0.69%
FH	84	0.40%
RS	99	0.47%
Others	19,163	91.35%
Total	20,977	100.00%

Table 3.1: Sample count and distribution across different categories.

Classification Model

We use the pre-trained BERT-Base (uncased) model (Devlin et al., 2018) and fine-tune it for multi-class text classification. The base model has 12 transformer blocks (i.e., hidden layers), a hidden size of 768, 12 attention heads, and 110 million parameters (Devlin et al., 2018). The model is pre-trained for English on uncased Wikipedia and BooksCorpus. For fine-tuning the model, we use *Adam* optimizer with a learning rate of $2e-5$, $\epsilon = 1e-8$,

L2 weight decay of 0.01, learning rate warmup over the first 500 steps with linear decay and Cross-Entropy Loss function. We observe the learning curve and find 5 epochs as optimal. Any example longer than the 512 tokens is truncated to comply with the token length restriction. For each example, we apply softmax over the logits returned by the model and pick the label with the highest probability.

NLG Module

In our previous work (Kazi and Kahanda, 2019), we develop an NLG module that handles all types of sentences except for short answer QA (Question-Answer) pairs where the answer alone does not provide the full context. In such cases, we need to mine information from both the question and the answer to formulate an assertive sentence that we can put in the clinical note. These QA pairs can cover a wide variety of question types and writing a rule for each type is time-consuming and is not an efficient solution. After examining clinical notes of behavioral health clinicians, we discover a simple way of formalizing these QA pairs. In our improved model, we formalize these QA pairs by prefixing the question with "when asked" and the answer with "he replied" and by joining the question and answer parts with a single comma as shown in Table 3.2.

#	Dialogue turn	Formal form
1	Provider: Where do you work? Patient: Near the mall.	When asked "where do you work", he replied "near the mall".
2	Provider: When did you wake up last night? Patient: It was before 4.	When asked "when did you wake up last night", he replied "it was before 4".
3	Provider: How does he push her? Patient: Not like with hands, just ignores her to make her mad.	When asked "how does he push her", he replied "not like with hands, just ignores her to make her mad".

Table 3.2: Text formalization of short answer questions.

Experimental Setup

We evaluate the performance of our classification model using 5-fold stratified cross-validation. While fine-tuning the BERT model, we feed the model three examples at a time due to GPU memory limitation. We report the performance of our model using Precision, Recall, and F1 scores. Since the category Others represent irrelevant information, we exclude it from the performance metrics.

Results and Discussion

In our previous preliminary work, we find SVMs with linear kernel performs the best with a macro-average AUROC score of 0.76. However, the dataset consisted of only 18 transcripts whereas 13 of them were synthetic. When we apply the same model over the new dataset, not a single sample gets labeled with an EHR category (data not shown). As we can see from the Table 3.1, only 8.65% of the samples are labeled with EHR categories (whereas in our previous work it was 89.66%). There are many samples that can be labeled with at least one EHR category but are labeled with *Others* because the information in these samples is not important enough to be included in the clinical note. The previous SVM model suffers to distinguish between the important and unimportant samples and predicts all the samples as unimportant due to its majority in the dataset. We also apply Decision Tree, Random Forests, and Multi-layer Perceptron classifiers (data not shown) and all of them have displayed performance similar to the SVM model.

In this project, we fine-tune a BERT model as a multi-class classification problem and assess the performance of this model. Comparatively, the BERT model performs significantly better than the SVM model on extracting important information. The performance of the BERT model is shown in Table 3.3 and Figure 3.3. The BERT model achieves an F1 score of 0.94 for the *Others* category which means the model can distinguish between the

Category	Precision	Recall	F1
CD	0.14	0.09	0.11
CC	0.24	0.16	0.19
HPI	0.09	0.08	0.09
PPH	1.00	0.08	0.15
HSU	1.00	0.33	0.50
SH	0.15	0.13	0.14
FH	0.06	0.06	0.06
RS	0.33	0.05	0.09
Others	0.93	0.96	0.94
Macro	0.44	0.22	0.25
EHRC Macro	0.38	0.12	0.18

Table 3.3: Performance of the BERT classification model. *Macro* refers to the overall performance of the model and *EHRC Macro* refers to the classification performance over only EHR Categories.

important and unimportant samples to a great extent. However, the model shows scope for improvement in classifying the important information. The BERT model achieves an overall F1 score of 0.18 on EHR categories. Among the EHR categories, the model performs the best for HSU with an F1 score of 0.50 and the worst for FH with an F1 score of 0.06. Since we are using the labels predicted by the classification model for generating clinical notes, it is more important to achieve higher precision than recall. Because, a reader would prefer an incomplete note over a note with unorganized information, e.g. half of the PPH section includes information on HPI. Considering that, the BERT model performs significantly better for PPH (P=1.00) than CC (P=0.24) even though the F1 score of CC (0.19) is higher than PPH (0.15).

We find no correlation between the sample counts and the performance. HSU has the highest F1 score with the lowest sample count. CC and PPH have similar F1 scores (0.19 and 0.15, respectively) whereas the sample count of CC is more than four times greater than PPH (564 and 121, respectively).

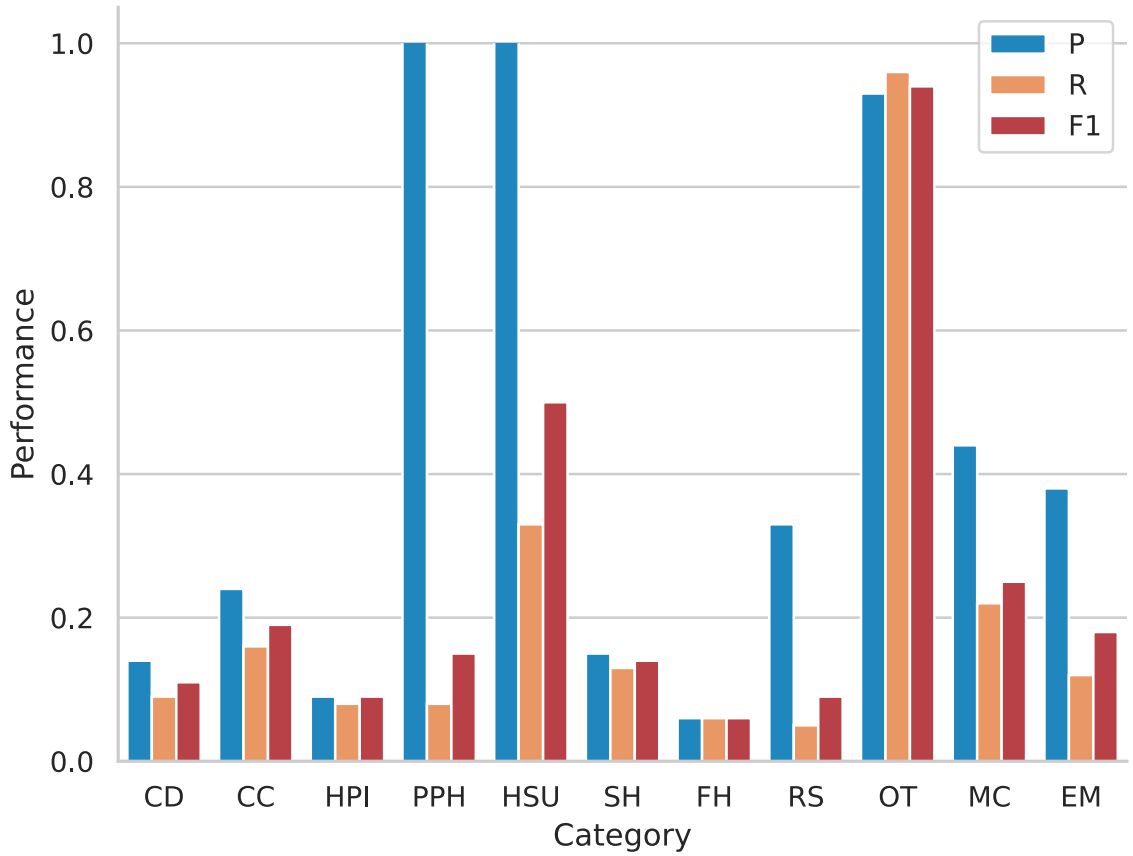

Figure 3.3: Performance of the BERT classification model. *MC* refers to the overall performance of the model and *EM* refers to the classification performance over only EHR Categories.

We try a couple of different strategies to further improve the performance but none of them is able to achieve a better score (data not shown). We divide the classification task between two BERT models, one to filter out the important samples (labeled with EHR categories) and another to classify the samples returned by the first model, but they are not able to outperform our current single model performance. We also fine-tune single BERT models with a portion of *Others* samples but keeping all the samples of all EHR categories (i.e. downsampling *Others* category). We use 5K, 1K, and 100 samples of *Others* (re-sampling at random) and in each case, we observe a decrease in overall performance.

We add the last rule to complete our rule-based NLG module. Sometimes providers note down the patient's answer as given preserving the wording of the answer. We adopt this strategy to develop the last rule of our model. Sentences generated using this rule will make a clinical note sounds a little technical but it preserves the wording of the answer as given by the patient which sometimes is very important in a clinical note. All short answer questions are some form of wh-questions. We can use part-of-speech tagging over the question to identify the subject, verb as well as wh-part and then replace the wh-part with the answer. This strategy works for some cases (e.g. Table 3.2, example 1) but not for all (e.g. Table 3.2, examples 2-3). Extracting the exact answer in all cases using rules is too expensive, if not impossible, considering the infinite possibilities in answers. In this regard, our proposed rule elegantly handles all cases.

Psychiatrists usually spend 20-30 minutes in documentation after a 45-60 minutes long therapy session (Miller, 2015). However, the average time to read a clinical note by a physician is 112 seconds and the reading time increases linearly with the character count (Brown et al., 2014). The longest time, 121 seconds, is recorded for the note with the highest character count. Usually, a conversation between a psychiatrist and a patient is very speech intensive and we can assume that it will result in a long clinical note but should not take more than 180 seconds to review a draft automatically generated from the conversation. We can also assume that the draft is missing some important information and includes some miscategorized or unimportant information as well. In our design, a user can drag and drop a single sentence or a chuck/ group of sentences from the transcript to the note and the NLG module takes care of the formalization automatically. This eliminates the need of updating the note by typing and will likely take the same amount of time to update the note as it takes to review it. Thus, we estimate the time to be about 6 minutes on average a behavioral clinician would spend on documentation when an automated draft is provided. This reduces the documentation time by 70-80%. It is believed that the use of

deep learning and natural language processing will reduce the documentation time by 75% (Matt Kuntz, the Executive Director of NAMI - Montana; personal communication).

Conclusion and Future Work

In this project, we focus on developing a pipeline that generates clinical notes directly from audio recordings of patient-provider conversations. We produce audio recordings that closely imitate real patient-provider conversations. Then, we transcribe the audio recordings using Google Cloud Platform Speech-to-Text API and compile a training dataset for the classification task. We fine-tune a BERT-base model to classify the samples with eight EHR categories. The BERT model shows its potential with the scope of further improvement for the classification task. On the text formalization task, we complete our previously developed rule-based model by adding a simple rule that elegantly handles the sophisticated case. Overall, we show the viability and potential of our pipeline in generating clinical notes directly from the patient-provider conversations.

In the future, we would like to investigate the model performance of XLNet (Yang et al., 2019) and RoBERTa (Liu et al., 2019) as classification models to extract important information from the transcripts. Both XLNet and RoBERTa have shown improved performance over BERT in most NLP tasks. The NLG module can be replaced with a Text2Text generation model, e.g. Phrase-BERT (Wang et al., 2021), which will downstream the text formalization process and will create a means of customizing the generated text with custom phrases and writing styles. However, fine-tuning a Text2Text model will require a larger dataset with at least 300 transcripts. We also plan to mix an equal number of machine generated notes with human-written notes and get them evaluated by domain experts using PDQI-9 (Stetson et al., 2012) to access the quality of these auto-generated notes. If an opportunity arises, we would also like to compare the quality of our notes with notes generated by DeepScibe.

CONCLUSION

Clinical documentation is one of the cornerstones of clinical practice and plays a major role in establishing evidence-based patient care and ensuring clinicians are reimbursed for provided care. Most clinicians spend 51.75% of their working hours in EHR including 12.26% in clinical documentation alone (Arndt et al., 2017). Time-consuming data entry is one of the major problems of EHRs (Payne et al., 2015), and sometimes clinicians spend about twice the time in clinical documentation than on direct patient care (Sinsky et al., 2016). Behavioral health clinicians collect most of their patient information over speech-intensive conversations, and residents have reported spending 22 hours per week in clinical documentation leading to clinician burnout (Domaney et al., 2018).

In this feasibility study, we develop a pipeline consisting of multiple machine learning models to fully automate the clinical documentation process from patient-provider conversations for behavioral health clinicians. We divide the overall problem into several sub-tasks and mainly focus on two sub-tasks, extraction as well as categorization of important information and text formalization while using existing tools/ software for the rest of the pipeline.

We start by developing a dataset of 18 transcripts composed of mostly synthetic transcripts and annotate them with five EHR categories. We train an SVM model and it shows promising results on categorizing information from digital transcripts. Later, we discover some limitations and downsides of using synthetic and human transcribed transcripts for automation. So, we put our effort into generating 65 audio recordings that closely imitates real patient-provider conversation and transcribe them using the state-of-the-art SR API. Then, we compose a new dataset using these SR transcribed transcripts annotated with eight EHR categories. We re-train the SVM model over this new dataset and observe a surprising drop in performance. After trying a few other machine learning models without

success, we divert our path to transformer-based deep learning models, more specifically, BERT models. We train a BERT-base model and it shows significantly high performance over other models on extracting important information from transcripts and categorizing them with EHR categories.

For generating a formal clinical note from a casual conversation, we develop a rule-based NLG module that takes informal sentences as well as QA pairs and converts them into assertive sentences in formal form. In our first attempt, we include rules to handle all types of inputs except for short-answer questions which require sophisticated rules to extract information from both the question and the answer to construct an assertive sentence in formal form. Later, we find a rather simple and easy way of handling short-answer questions upon exploring clinical notes. We add a new rule for short-answer questions and complete our NLG module.

Our pipeline uses existing tools/ software to record a patient-provider conversation, SR technologies to transcribe audio recordings into digital transcripts, the BERT model to extract and categorize important information, and the NLG module to compose a formal note from the extracted information. The pipeline provides a draft clinical note that eliminates the need of writing a clinical note from scratch which is very time-consuming. Our automated process has the potential of reducing the clinical documentation burden by 70-80% for behavioral health clinicians. Mental health is one of the most expensive and largest public healthcare burdens in the United States (Soni, 2009; Diederich and Song, 2014). With reduced documentation load, providers can allocate more time to direct patient care, which will not only increase providers' availability but also will make mental healthcare less expensive.

While our results are very promising, we identify several avenues for future research. A new classification model can be developed using XLNet (Yang et al., 2019) or RoBERTa (Liu et al., 2019). Both models have out-performed BERT (Devlin et al., 2018) on most NLP

tasks and have the potential of showing better performance in information extraction and classification. However, developing a new classification model with either XLNet or RoBERTa will require a larger dataset of at least 200 transcripts. The NLG module can also be improved by using a language model like Phrase-BERT (Wang et al., 2021) that has the potential of automating the text formalization process and personalizing the NLG module by adapting to the clinician's writing styles and custom phrases. The development of a deep learning-based NLG module will also require a larger dataset of at least 300 transcripts. We plan to evaluate the quality of our clinical notes using PDQI-9 (Stetson et al., 2012). If an opportunity arises, we would also like to compare our clinical notes with notes generated by DeepScribe.

REFERENCES CITED

Mohammad Anani, Nazmul Kazi, Matthew Kuntz, and Indika Kahanda. 2019. RDoC task at BioNLP-OST 2019. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 216–226, Hong Kong, China. Association for Computational Linguistics.

Brian G Arndt, John W Beasley, Michelle D Watkinson, Jonathan L Temte, Wen-Jan Tuan, Christine A Sinsky, and Valerie J Gilchrist. 2017. Tethered to the ehr: Primary care physician workload assessment using ehr event log data and time-motion observations. *The Annals of Family Medicine*, 15(5):419–426.

Jim Atherton. 2011. Development of the electronic health record. *AMA Journal of Ethics*, 13(3):186–189.

Ashley K Barrett. 2018. Electronic health record (EHR) organizational change: Explaining resistance through profession, organizational experience, and ehr communication quality. *Health communication*, 33(4):496–506.

David W Bates, Jonathan M Teich, Joshua Lee, Diane Seger, Gilad J Kuperman, Nell Ma'Luf, Deborah Boyle, and Lucian Leape. 1999. The impact of computerized physician order entry on medication error prevention. *Journal of the American Medical Informatics Association*, 6(4):313–321.

James P Becker, Indika Kahanda, and Nazmul H Kazi. 2021. WIP: Detection of student misconceptions of electrical circuit concepts in a short answer question using NLP. In *2021 ASEE Virtual Annual Conference Content Access*.

Viv Bewick, Liz Cheek, and Jonathan Ball. 2004. Statistics review 13: receiver operating characteristic curves. *Critical care*, 8(6):508.

Christopher M Bishop. 2006. *Pattern recognition and machine learning*. Springer.

Suzanne V Blackley, Valerie D Schubert, Foster R Goss, Wasim Al Assad, Pamela M Garabedian, and Li Zhou. 2020. Physician use of speech recognition versus typing in clinical documentation: A controlled observational study. *International Journal of Medical Informatics*, 141:104178.

British Council. 2019. Questions and negatives. *Learn English British Council, retrived from: https://learnenglish.britishcouncil.org/en/english-grammar/questions-and-negatives*.

PJ Brown, JL Marquard, B Amster, M Romoser, J Friderici, S Goff, and D Fisher. 2014. What do physicians read (and ignore) in electronic progress notes? *Applied clinical informatics*, 5(02):430–444.

Bruce G Buchanan, Johanna D Moore, Diana E Forsythe, Giuseppe Carenini, Stellan Ohlsson, and Gordon Banks. 1995. An intelligent interactive system for delivering individualized information to patients. *Artificial intelligence in medicine*, 7(2):117–154.

Dean Carcone and Anthony C Ruocco. 2017. Six years of research on the National Institute of Mental Health's Research Domain Criteria (RDoC) initiative: a systematic review. *Frontiers in cellular neuroscience*, 11:46.

Joan A Casey, Brian S Schwartz, Walter F Stewart, and Nancy E Adler. 2016. Using electronic health records for population health research: A review of methods and applications. *Annual review of public health*, 37:61–81.

Hongshen Chen, Zhaochun Ren, Jiliang Tang, Yihong Eric Zhao, and Dawei Yin. 2018. Hierarchical variational memory network for dialogue generation. In *Proceedings of the 2018 World Wide Web Conference*, pages 1653–1662.

Robin A Cohen, Emily P Terlizzi, Amy E Cha, and Michael E Martinez. 2021. Health insurance coverage: Early release of estimates from the national health interview survey, january–june 2020. *National Center for Health Statistics*.

Tammy Combs. 2020. The importance of high-quality clinical documentation across the healthcare continuum. *Journal of AHIMA*.

Mario Cruz, Debra Roter, Robyn Flaum Cruz, Melissa Wieland, Lisa A Cooper, Susan Larson, and Harold Alan Pincus. 2011. Psychiatrist-patient verbal and nonverbal communications during split-treatment appointments. *Psychiatric Services*, 62(11):1361–1368.

Bruce N Cuthbert and Thomas R Insel. 2013. Toward the future of psychiatric diagnosis: The seven pillars of RDoC. *BMC medicine*, 11(1):126.

Deepscribe Inc. 2021. Medical prodigy deepscribe lands $5.2m seed funding for life-changing ai assistant.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Joachim Diederich and Insu Song. 2014. *Mental Health Informatics: Current Approaches*, pages 1–16. Springer Berlin Heidelberg, Berlin, Heidelberg.

Nicholas M Domaney, John Torous, and William E Greenberg. 2018. Exploring the association between electronic health record use and burnout among psychiatry residents and faculty: A pilot survey study. *Academic Psychiatry*, 42(5):648–652.

Eli Goldberg, Norbert Driedger, and Richard I Kittredge. 1994. Using natural-language processing to produce weather forecasts. *IEEE Expert*, 9(2):45–53.

Foster R Goss, Suzanne V Blackley, Carlos A Ortega, Leigh T Kowalski, Adam B Landman, Chen-Tan Lin, Marie Meteer, Samantha Bakes, Stephen C Gradwohl, David W Bates, et al. 2019. A clinician survey of using speech recognition for clinical documentation in the electronic health record. *International journal of medical informatics*, 130:103938.

J Henry, Yuriy Pylypchuk, Talisha Searcy, and Vaishali Patel. 2016. Adoption of electronic health record systems among us non-federal acute care hospitals: 2008–2015. *ONC data brief*, 35:1–9.

Hsiao and Hing. 2014. Use and characteristics of electronic health record systems among office-based physician practices: United states, 2001–2013. *NCHS Data Brief, No 143. Hyattsville, MD: National Center for Health Statistics.*

International Society for Biocuration. 2018. Biocuration: Distilling data into knowledge. *PLoS Biology*, 16(4):e2002846.

Lidija Iordanskaja, Myunghee Kim, Richard Kittredge, Benoit Lavoie, and Alain Polguere. 1992. Generation of extended bilingual statistical reports. In *COLING 1992 Volume 3: The 14th International Conference on Computational Linguistics*.

David R Kaufman, Barbara Sheehan, Peter Stetson, Ashish R Bhatt, Adele I Field, Chirag Patel, and James Mark Maisel. 2016. Natural language processing-enabled and conventional data capture methods for input to electronic health records: A comparative usability study. *JMIR medical informatics*, 4(4).

Eda Kavlakoglu. 2020. NLP vs. NLU vs. NLG: The differences between three Natural Language Processing Concepts.

Nazmul Kazi and Indika Kahanda. 2019. Automatically generating psychiatric case notes from digital transcripts of doctor-patient conversations. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 140–148.

Nazmul Kazi, Nathaniel Lane, and Indika Kahanda. 2021. Automatically cataloging scholarly articles using library of congress subject headings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 43–49, Online. Association for Computational Linguistics.

Jennifer King, Vaishali Patel, Eric W Jamoom, and Michael F Furukawa. 2014. Clinical benefits of electronic health record use: National findings. *Health services research*, 49(1pt2):392–404.

Sherry C Kwon and Abdolreza Saadabadi. 2021. *Mental Health Challenges In Caring For American Indians and Alaska Natives*. StatPearls Publishing, Treasure Island (FL).

Ronilda C Lacson, Regina Barzilay, and William J Long. 2006. Automatic analysis of medical dialogue in the home hemodialysis domain: Structure induction and summarization. *Journal of biomedical informatics*, 39(5):541–555.

Scott H Lee. 2018. Natural language generation for electronic health records. *NPJ digital medicine*, 1(1):1–7.

Scott O Lilienfeld and Michael T Treadway. 2016. Clashing diagnostic approaches: DSM-ICD versus RDoC. *Annual review of clinical psychology*, 12:435–463.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Dinah Miller. 2015. Addressing the shortage of psychiatrists: What keeps us from seeing more patients? *Clinical Psychiatry News*.

Miroslav Nagy, Petr Hanzlicek, Jana Zvarova, Tatjana Dostalova, Michaela Seydlova, Radim Hippman, Lubos Smidl, Jan Trmal, and Josef Psutka. 2008. Voice-controlled data entry in dental electronic health record. *Studies in health technology and informatics*, 136:529.

Nuance Communications, Inc. 2021. EPR Software: Dragon Medical Practice Edition.

J Marc Overhage and David McCallie Jr. 2020. Physician time spent using the electronic health record during outpatient encounters: A descriptive study. *Annals of internal medicine*, 172(3):169–174.

Thomas H Payne, Sarah Corley, Theresa A Cullen, Tejal K Gandhi, Linda Harrington, Gilad J Kuperman, John E Mattison, David P McCallie, Clement J McDonald, Paul C Tang, et al. 2015. Report of the AMIA EHR-2020 Task Force on the status and future direction of EHRs. *Journal of the American Medical Informatics Association*, 22(5):1102–1110.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

François Portet, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer, and Cindy Sykes. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7-8):789–816.

M Reinert, T Nguyen, and D Fritze. 2020. The state of mental health in america 2020. *Mental Health America, Alexandria VA*, 500:22314–1520.

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

Cynthia Reyes, Alissa Greenbaum, Catherine Porto, and John C Russell. 2017. Implementation of a clinical documentation improvement curriculum improves quality metrics and hospital charges in an academic surgery department. *Journal of the American College of Surgeons*, 224(3):301–309.

Michael Rigby, Jan Lindmark, and Pier Maria Furlan. 1998. The importance of developing an informatics framework for mental health. *Health Policy*, 45(1):57–67.

William J Rudman, John S Eberhardt, William Pierce, and Susan Hart-Hester. 2009. Healthcare fraud and abuse. *Perspectives in Health Information Management/AHIMA, American Health Information Management Association*, 6(Fall).

Christine Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. 2016. Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties. *Annals of internal medicine*, 165(11):753–760.

Anita Soni. 2009. The five most costly conditions, 1996 and 2006: Estimates for the us civilian noninstitutionalized population. statistical brief# 248. july 2009. *Rockville, MD: Agency for Healthcare Research and Quality*.

Srinivasan Sridhar, Nazmul Kazi, Indika Kahanda, and Bernadette McCrory. 2021. Psychiatry transcript annotation: Process study and improvements. *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*, 10(1):71–75.

Peter D Stetson, Suzanne Bakken, Jesse O Wrenn, and Eugenia L Siegler. 2012. Assessing electronic note quality using the physician documentation quality instrument (PDQI-9). *Applied clinical informatics*, 3(2):164.

Minal Thakkar and Diane C Davis. 2006. Risks, barriers, and benefits of EHR systems: A comparative study based on size of hospital. *Perspectives in Health Information Management/AHIMA, American Health Information Management Association*, 3.

Adele L Towers. 2013. Clinical documentation Improvement - A physician perspective: Insider tips for getting physician participation in CDI programs. *Journal of AHIMA*, 84(7):34–41.

Donna Vanderpool. 2012. HIPAA - Should I Be Worried? *Innovations in clinical neuroscience*, 9(11-12):51.

Markus Vogel, Wolfgang Kaisers, Ralf Wassmuth, and Ertan Mayatepek. 2015. Analysis of documentation speed using web-based medical speech recognition technology: Randomized controlled trial. *Journal of medical Internet research*, 17(11):e5072.

Byron C Wallace, M Barton Laws, Kevin Small, Ira B Wilson, and Thomas A Trikalinos. 2014. Automatically annotating topics in transcripts of patient-provider interactions via machine learning. *Medical Decision Making*, 34(4):503–512.

Shufan Wang, Laure Thompson, and Mohit Iyyer. 2021. Phrase-BERT: Improved phrase embeddings from BERT with an application to corpus exploration. In *Proceedings of the*

2021 Conference on Empirical Methods in Natural Language Processing, pages 10837–10851, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Seena Zierler-Brown, Timothy R Brown, David Chen, and Robert Wayne Blackburn. 2007. Clinical documentation for patient care: Models, concepts, and liability considerations for pharmacists. *American Journal of Health-System Pharmacy*, 64(17):1851–1858.

APPENDICES

APPENDIX A

RDOC TASK AT BIONLP-OST 2019: A MENTAL HEALTH INFORMATICS TASK WITH RESEARCH DOMAIN CRITERIA

RDoC Task at BioNLP-OST 2019: A Mental Health Informatics Task with Research Domain Criteria

Mohammad Anani[1*], **Nazmul Kazi**[1*], **Matthew Kuntz**[23] **and Indika Kahanda**[1]
[1] Gianforte School of Computing, Montana State University, MT, USA
[2] National Alliance of Mental Illness (NAMI) Montana, Helena, MT, USA
[3] Center for Mental Health Research and Recovery, Montana State University, MT, USA
mohammad.anani@student.montana.edu
{kazinazmul.hasan,matthew.kuntz,indika.kahanda}@montana.edu
[*] The authors wish it to be known that Mohammad Anani and Nazmul Kazi should be regarded as joint first authors.

Abstract

BioNLP Open Shared Tasks (BioNLP-OST) is an international competition organized to facilitate development and sharing of computational tasks of biomedical text mining and solutions to them. For BioNLP-OST 2019, we introduced a new mental health informatics task called "RDoC Task", which is composed of two subtasks: information retrieval and sentence extraction through National Institutes of Mental Health's Research Domain Criteria framework. Five and four teams around the world participated in the two tasks, respectively. According to the performance on the two tasks, we observe that there is room for improvement for text mining on brain research and mental illness.

1 Introduction and Motivation

The breadth of brain research is too expansive to be effectively curated without computational tools especially involving machine learning models. For example, a Pubmed search for "Brain" on August 12, 2019, revealed 854,612 articles[1]. More specifically, an August 12, 2019 search for the single mental illness diagnosis of "depression" revealed 530,519 articles[2]. And a search for anxiety revealed 224,305 articles[3]. It is not possible for researchers to functionally analyze all of the critical data patterns both within a single diagnosis or across diagnoses that could be revealed by those articles.

The challenge of curating brain research has been further complicated by the National Institute of Mental Health's adoption of the Research Domain Criteria (RDoC) [6]. Since 1952, the Diagnostic and Statistical Manual of Mental Disorders and International Classification of Diseases [5] (popularly known as DSM and ICD, respectively), have "reigned supreme" as the single "overarching model of psychiatric classification" [14]. That supremacy began to crumble in 2010 when the National Institute of Mental Health launched the RDoC initiative, an alternate framework to conceptually organize and direct biological research on mental disorders [1]. The RDoC initiative intends "to foster integration not only of psychological and biological measures but also of the psychological and biological constructs those measures measure" [13].

The RDoC initiative has fostered significant debate among brain health researchers. It has also created a significant categorization challenge - specifically how to curate articles completed under the DSM-ICD criteria so their data can be incorporated into the RDoC model. Brain science cannot afford to lose critical insights from the numerous articles on different sides of the categorization divide. Hence, it is vital that all existing and future biomedical literature related to brain research is correctly categorized with respect to the RDoC terminology in addition to DSM-ICD models.

However, manual curation of brain research articles using RDoC terminology by human annotators can be highly resource-consuming due to several reasons. RDoC framework is comprehensive and complex. It is made up six major *domains* of human functioning, which is further broken down to multiple *constructs* that comprise different aspects of the overall range of functions[4]. The RDoC matrix helps describe these constructs using several *units of analysis* such as molecules and circuits. On top of this, the rate of publication of biomedical literature (and by extension brain re-

[1] Pubmed search for Brain conducted on August 12, 2019
[2] Pubmed search for depression conducted on August 12, 2019
[3] Pubmed search for anxiety conducted on August 12, 2019
[4] https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/definitions-of-the-rdoc-domains-and-constructs.shtml

Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, pages 216–226
Hong Kong, China, November 4, 2019.

search related literature) is growing at an exponential rate [10]. This means that the gap between annotated versus unannotated articles will continue to grow at an alarming rate unless more efficient means of automated annotation is developed soon.

In order to invite text mining teams around the world to develop informatics models for RDoC, we introduced the RDoC Task[5] at this years' BioNLP-OST 2019 workshop[6]. RDoC task is a combination of two subtasks focusing on a subset of RDoC constructs: (a) Task 1 (RDoC-IR) - retrieving PubMed Abstracts related to RDoC constructs, and (b) Task 2 (RDoC-SE) - extracting the most relevant sentence for a given RDoC construct from a known relevant abstract. Both these tasks represent two very important steps of the typical triage process [10], which are finding the articles related to RDoC constructs and then extracting a specific snippet of information that is useful for curation or downstream tasks such as automatic text summarization [15].

There have been several shared tasks on text mining from biomedical literature and clinical notes in the last decade [19, 12] as well as a few shared tasks related to mental health topics ([4, 18, 22, 21, 30]). CLPsych 2015 Shared Task [4] focused on identifying depression and PTSD users from twitter data, while the same task from the following year (i.e. CLPsych 2016 Shared Task [18]) revolved around classifying the severity of peer support forum posts. One of the i2b2[7] challenges from 2011 focused on the sentiment analysis of suicide notes [22, 21].

In 2017, Uzuner et al. introduced the "The RDoC for Psychiatry" challenge, which was composed of three tracks: de-identification of mental health records [28], determination of symptom severity from a psychiatric evaluation of a patient) related to one of the RDoC domains) [9], and the use of mental health records released through the challenge for answering novel questions [32, 29, 7]. In contrast, the RDoC task is a combination of information retrieval and sentence extraction from Biomedical literature related to RDoC constructs.

To generate benchmark data for the RDoC task, three annotators were used to curate the gold-standard datasets. The registration for the RDoC Task opened in March of 2019. Over 30 teams around the world registered for the two tasks. Training data in two batches were released in the month of April. Test data, again in two batches, were released in June. The participants were asked to submit their final predictions by June 19. Eventually, 4 and 5 groups each competed in Tasks 1 and 2, respectively. The final results were made public immediately after the submission deadline.

Two (out of four) and four (out of five) teams each outperformed the baseline methods in task 1 and 2, respectively. The increase in performance over the baselines were more noticeable in task 2 suggesting that information retrieval for RDoC task may be more challenging. There was quite a lot of variation across the several RDoC constructs used for the tasks suggesting that the complexity of different constructs may hinder certain models and construct-specific methods or models may be a requirement in the future. Overall observations from the RDoC Task highlights the need for more sophisticated method development.

The rest of the paper is organized as follows. Section 2 describes the benchmark or gold-standard data preparation process, development of training and test sets, submission requirements, baseline methods used by the organizers, and the performance measures used for the evaluation. Section 3 presents and discusses the overall results for the two tasks. Finally, Section 4 summarizes the task findings as well as describes the potential future work.

2 RDoC Task setup

RDoC Task is a combination of two subtasks. Participants were allowed to choose to participate in one or both tasks. Task 1 is on retrieving PubMed Abstracts related to RDoC constructs, while Task 2 is on extracting the most relevant sentences for an RDoC construct from an already relevant abstract.

In task 1, participants are given a set of PubMed abstracts and they are required to rank abstracts according to relevance for various RDoC constructs. In task 2, participants are given a set of PubMed abstracts relevant for an RDoC construct, and they are required to extract the most relevant sentence from each abstract for the corresponding RDoC construct.

[5] https://sites.google.com/view/rdoc-task/home
[6] http://2019.bionlp-ost.org
[7] https://www.i2b2.org/

2.1 Timeline

The RDoC Task was organized in two main phases (a) *Training* phase (8 weeks, from April-June 2019), and (b) *Evaluation* phase (1 week in mid-June). At the beginning of the training phase, participants were provided with labeled data (i.e. Training data) and they were expected to develop and fine-tune their models using these known labels. At the beginning of the Evaluation phase, unlabeled data (i.e. Test data) was made available to the participants. They were required to predict labels for this data and submit the predictions to the organizers at the end of the Evaluation phase. Finally, the organizers used the (with-held) labels of the test data for evaluating the accuracy of submissions.

2.2 The benchmark preparation

For the RDoC Task, 8 RDoC constructs out of 25 total constructs from the latest version of the RDoC matrix[8] were used. The motivation was to restrict ourselves to a subset of RDoC framework for which benchmark data can be gathered within a reasonable time-frame. However, these 8 constructs completely cover two of the six domains in the RDoC framework – namely *Negative Valence Systems* and *Arousal and Regulatory* Systems as shown in Table 1.

Table 1: Subset of RDoC constructs used for this task and their domain.

Domain	Construct
Negative Valence Systems	Acute Threat (Fear) Potential Threat (Anxiety) Frustrative Nonreward Sustained Threat Loss
Arousal/Regulatory Systems	Arousal Circadian Rhythms Sleep and Wakefulness

Under the guidance of the Subject Matter Experts from the National Alliance of Mental Illness (NAMI) Montana, the RDoC task benchmark was created by using Entrez e-search utility [26] to search the PubMed database to collect abstracts related to RDoC constructs. That is, we start by using the RDoC construct name as the only keyword to retrieve relevant articles.

If such an approach does not generate the desired number of articles or is too ambiguous on its own (e.g., *Loss* construct), we have utilized terms from the *Behaviors* unit of the RDoC matrix in addition to the construct name.

For example, the The query used for *Loss* construct was "Loss""Amotivation" or "Loss""Anhedonia" or "Loss""Crying" or "Loss""Guilt" or "Loss""Rumination" or "Loss""Sadness" or "Loss""Shame" or "Loss""Withdrawal" or "Loss""Worry". This retrieves about 315 articles, whereas using only "Loss" as the sole query retrieves too many articles (approximately one million articles).

Other queries follow a similar format as Loss when very few (<200) or too many (>10,000) articles were retrieved with the RDoC construct name as the only keyword. 200 abstracts was the desired minimum number of abstracts per construct that we were planning to send to each annotator. So, if the initial search retrieved less articles, it was deemed too narrow for our objective, and we added terms from the *Behavior* elements belonging to that construct to retrieve more than 200 articles. For example, for the construct Frustrative Nonreward, a PubMed search with the construct name only returns 52 abstracts (retrieved on 09/30/2019)[9]. The RDoC page for Frustrative Nonreward contains one element under the Behavior unit: "physical and relational aggression"[10]. Then, using this term, the search query becomes: "Frustrative Nonreward" or "physical aggression" or "relational aggression", which returns 736 abstracts.

10,000 was a rough estimation of an excessively inclusive search term as determined by our Subject Matter Expert. In other words, the construct name on its own (construct Loss, for example) has a very general definition, resulting in retrieving a large heterogeneous set of articles. Therefore, in these situations, other more specific terms describing the construct were used to limit the scope. Upon generating a search query that retrieves a satisfactory number of articles, we sort them by relevance to the query used.

Then the above-retrieved articles were provided

[8] https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/constructs/rdoc-matrix.shtml

[9] https://www.ncbi.nlm.nih.gov/pubmed/?term=Frustrative+Nonreward

[10] https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/constructs/frustrative-nonreward.shtml

to three annotators for curation (an example of the annotation guidelines used is available online[11]). For each construct, they were asked to read the title and the abstract and determine whether it provides enough evidence that the abstract was related to the construct. If it was related it was annotated as "positive" (or "negative" otherwise). In addition, they were asked to identify up to 3 most relevant sentences to the abstract (i.e. the sentences that provide most evidence that the abstract is related to the said construct). The inter-annotator agreements are given in Table 2. Example annotation of an abstract is depicted in Figure 1.

While acknowledging we generated a *closed* set of articles for the information retrieval task, we emphasize that this complete process was guided by NAMI experts. They typically use keyword search for first finding relevant articles. Then they use manual curation to remove false positives. Hence, our benchmark datasets are developed using this approach. We wanted the RDoC Task to resemble how a typical curator would find information in this domain.

Table 2: Inter-annotator agreement of Task 1 and Task 2. κ_{free}: Free-Marginal Multirater Kappa [24] computed online[12]

RDoC Construct	κ_{free} Task 1	κ_{free} Task 2
Acute Threat	0.37	0.24
Potential Threat	0.45	0.27
Frustrative Nonreward	0.24	0.20
Sustained Threat	0.18	0.14
Loss	0.25	0.29
Arousal	0.64	0.35
Circadian Rhythms	0.95	0.35
Sleep & Wakefulness	0.97	0.51

We consolidated the labels from the three annotators using the majority vote (i.e. if at least 2 annotators agreed on a label, that was used as the final label for the abstract). In addition, we collected all the most relevant sentences by the three annotators (i.e. set union) as the final set of sentences. This means each abstract could have up to 9 most relevant sentences. In our dataset, at most 6 sentences were observed. This consolidated data was used to create training and test sets as described below.

We believe that the task of identifying the most relevant sentence was more challenging for the annotators than the task of identifying whether a given abstract was related to an RDoC construct or not (for the latter task, annotators were choosing between two labels while for the former, they were choosing from k sentences in the abstract). Therefore, it was possible that there would be more variability in annotations for the former task. So, we used the set union to allow for more flexibility.

2.3 Train, Test and Submission data

In the context of the RDoC task, training data refers to the labeled data sets initially provided to the participants for developing their models. Test sets refer to the unlabeled (i.e. with withheld labels) data sets for which they were asked to submit predictions. All the datasets are available online[13].

For each construct, two separate sets of articles (referred to as Set 1 and Set 2) were annotated. Data from the Set 1 and Set 2 were allocated for training and test data, respectively. Annotators were not aware of this distinction. Set 1 and Set 2 splits were randomly performed per each construct separately before annotation. Therefore, explicit stratified sampling was not applicable.

For each construct, a random subset of positive examples from Set 1 was used as the training examples for both Task 1 and 2 (negative examples were not provided). 80% of random abstracts from Set 2 were used as the test set for Task 1 (this included both positive and negative examples). The subset of positive examples in the rest of the Set 2 (i.e. 20%) was used as the test set for Task 2 (negative examples were not used).

2.3.1 Train data

As mentioned above, we provided the participants of the RDoC task with training examples for each of these 8 RDoC constructs. For task 1, the training examples are randomly selected subsets of positive abstracts for each of the RDoC constructs as shown in Table 3. For task 2, we provided up to 6 most relevant sentences for each of the abstract provided as part of Task 1 train data. In other words, the same set of PubMed IDs were used for training data of both tasks. The distribution of the training examples across the eight constructs is

[11] https://montana.box.com/s/kh0hmyn1jcj5ajvr2nibq4iwwgiv3led

[13] https://www.cs.montana.edu/rdoc-task/data/

Title: Characteristics of Physical Aggression in Children of Immigrant Mothers and Non-immigrant Mothers: A Cross-Sectional Analysis of the Survey of Young Canadians.

Abstract: Physical aggression (PA) is important to regulate as early as the preschool years in order to ensure healthy development of children. This study aims to determine the prevalence and characteristics of PA in children of immigrant and non-immigrant mothers. Bivariate and multivariable logistic regression was performed, with the outcome, PA, and covariates including maternal, child, household and neighbourhood characteristics. Twenty percent of children of non-immigrant mothers and 16% of children of immigrant mothers reported PA. The characteristics of PA differ between children of immigrant versus non-immigrant mothers therefore healthcare providers, policy makers, and researchers should be mindful to address PA in these two groups separately, and find ways to tailor current recommended coping strategies and teach children alternative ways to solve problems based on their needs.

RDoC Construct: Sustained Threat

Figure 1: An example of annotating an abstract for both Task 1 and Task 2. The abstract is annotated positive for *Sustained Threat* (Task 1; highlighted in purple) and the most relevant sentence in the abstract is identified (Task 2; highlighted in yellow).

provided in the Table 3 and the distribution of the number of most relevant sentences per construct is shown in Table 4.

Table 3: The number of training examples (positively labeled abstracts) provided for Tasks 1 and 2 across constructs.

RDoC construct	# Abstracts	%
Acute Threat (Fear)	39	14.7
Potential Threat (Anxiety)	27	10.2
Frustrative Nonreward	21	7.9
Sustained Threat	18	6.8
Loss	28	10.5
Arousal	38	14.3
Circadian Rhythms	47	17.7
Sleep and Wakefulness	48	18.1
Total	266	100.0

2.3.2 Test data

The Task 1 test set provided the participants with a random list of 999 relevant (positive) and irrelevant articles (negative) for each of the RDoC constructs (but without the actual labels). The label distribution is given in Table 5. The task 2 test set provided the participants with a list of relevant articles from which they had to extract a relevant sentence with respect to the given RDoC category. The set of abstracts used for test sets of task 1 and 2 were mutually independent for obvious reasons. The distribution of the test set for task 2 across constructs is shown in Table 6 and the distribution of the number of most relevant sentences per construct is provided in Table 4.

2.3.3 Participant Submissions

For task 1, participants were required to submit scores for each abstract in the test set. Scores should correspond to the predicted relevance of the abstract to the given construct. For task 2, participants were required to submit sentences from each abstract that is predicted as the most relevant sentence to the given construct. Submitting a score was not required.

Participants uploaded their submissions through an online web application[14]. We designed the web system to validate the content format of each submission before uploading the file(s) in the server. Upon finding a line that is not properly formatted, the system alerts the participant with an error message including the ill-formatted line number. If the file(s) are properly formatted, the system uploads the submission in the server, automatically analyzes the submission using python scripts and immediately reports the scores of two selected constructs, *Acute Threat (Fear)* and *Loss*, back to the participant.

The participants were allowed to make an un-

[14]https://www.cs.montana.edu/rdoc-task/

Table 4: Distribution of the number of most relevant (gold-standard) sentences in abstracts for each construct in the training data. #x: the percentage of abstracts with x relevant sentences.

RDoC Construct	Train Data						Test Data			
	#1	#2	#3	#4	#5	#6	#1	#2	#3	#4
Acute Threat (Fear)	0.0	15.4	35.9	35.9	10.3	2.6	15.8	31.6	42.1	10.5
Potential Threat (Anxiety)	11.1	33.3	55.6	0.0	0.0	0.0	38.2	35.3	20.6	5.9
Frustrative Nonreward	4.8	47.6	47.6	0.0	0.0	0.0	54.3	37.1	8.6	0.0
Sustained Threat	5.6	61.1	33.3	0.0	0.0	0.0	38.9	41.7	16.7	2.8
Loss	10.7	25.0	42.9	21.4	0.0	0.0	61.8	32.4	5.9	0.0
Arousal	7.9	63.2	28.9	0.0	0.0	0.0	23.1	53.8	15.4	7.7
Circadian Rhythms	2.1	51.1	46.8	0.0	0.0	0.0	20.0	40.0	26.7	13.3
Sleep and Wakefulness	10.4	62.5	27.1	0.0	0.0	0.0	26.7	36.7	30.0	6.7

Table 5: The number of abstracts in test set for task 1. Pos and %: number of positively labeled abstracts and their percentages, and Neg: number of negatively labeled abstracts.

RDoC construct	# Pos	%	# Neg
Acute Threat (Fear)	53	67.1	26
Potential Threat (Anxiety)	124	89.2	15
Frustrative Nonreward	96	66.7	48
Sustained Threat	82	56.2	64
Loss	90	65.2	48
Arousal	97	89.8	11
Circadian Rhythms	123	100.0	0
Sleep and Wakefulness	121	99.2	1
Total	786	78.7	213

Table 6: The number of abstracts and their percentages in test set for task 2.

RDoC construct	# Abstracts	%
Acute Threat (Fear)	19	7.8
Potential Threat (Anxiety)	34	13.9
Frustrative Nonreward	35	14.3
Sustained Threat	36	14.8
Loss	34	13.9
Arousal	26	10.7
Circadian Rhythms	30	12.3
Sleep and Wakefulness	30	12.3
Total	244	100.0

limited number of submissions and the scores from past submissions were discarded upon a new submission. This meant they could re-submit until they achieved a satisfactory performance for the above two constructs. The performance scores for all the constructs were made available immediately after the submission deadline. The older scores were only discarded for the purposes of the final evaluation. However, these scores are retained for potential future research.

2.4 Baseline methods

We used TF-IDF [23] with smooth IDF weights and cosine similarity [27] to calculate the similarity score for each document against a query and used these scores to rank the documents by relevance. Regardless of the task, we used the corresponding construct name concatenated with its definition as the query string. We used the definitions of constructs as defined by the National Institute of Mental Health listed online[15].

For task 1, each document is the title concatenated with the corresponding abstract and the similarity scores are used to rank the articles for each construct. For task 2, documents are the sentences of the abstracts and the top-ranked sentence per abstract was returned based on the similarity scores. All the baseline models were implemented using the Scikit-learn Python library [20]. No preprocessing techniques were applied to the abstract text. In addition to the above TFIDF-based baseline, we also used BM25 [25] as a baseline. But due to its comparatively lower performance on both tasks 1 and 2, BM25 values are not reported in this paper.

[15] https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/definitions-of-the-rdoc-domains-and-constructs.shtml

2.5 Metrics used for evaluation

For task 1, we use Mean Average Precision (MAP) [16] as the performance measure because it is one of the most frequently used measures for IR [31, 8, 11]. First, we compute the Average Precision (AP) for each construct independently and macro-average across the constructs to compute the Mean Average Precision. For task 2, due to the non-applicability of utilizing popular standard measures such as precision and recall [3], we define the *Accuracy* as the percentage of abstracts with correctly predicted most relevant sentence. If at least one of the gold-standard sentences match the predicted sentence, it is counted as 1 and 0 otherwise (therefore, note that this measure is not the same as the typical accuracy measure used in Natural Language Processing and Machine Learning. We average across constructs to get the *Macro Average Accuracy*.

It should be pointed out that, technically, there is no "negative" class for the task 2 (in the traditional sense used for predictive models). Participants are given abstracts already known to be relevant to a construct. They are asked to submit just one sentence that they think is the most relevant (or that helps them the most for finding the relevance between the given abstract and the construct). Hence the participants are unable to gain undue advantages due to any class imbalances even though the above-defined performance measure may closely resemble the typical "Accuracy". Also, since we did not collect confidence scores for task 2, we did not compute threshold independent measures such as AUROC (area under the ROC curve).

3 Results and Discussion

Inter-annotator agreements for many of the constructs in both tasks 1 and 2 are relatively low (see table 2). According to the annotators, there were several reasons why information retrieval and sentence extraction with RDoC was reasonably challenging. The very generalized nature of the RDoC constructs, as well as ambiguity in the language stating the purpose/hypothesis/results of the experiment, made it difficult to find the relevance of a given abstract to an RDoC construct. The way the abstracts were written, made it seem such that it could be potentially tied to/or not, to various RDoC sentences.

Annotators reported that they had difficulties with the 'Sustained Threat' and the 'Frustrative Non-Reward' constructs. For example, some annotators felt that every abstract that they read was related to Frustrative Non-Reward construct because many of the abstracts specifically studied the relational and physical aggressive behaviors. Although a lot of the studies tested these behaviors, it was challenging to figure out if they were "directly" related to Frustrative Non-Reward or not. For instance, several studies comparatively tested relational and physical aggression between genders (2 behaviors of Frustrative Non-Reward), but the abstracts didn't explicitly mention "withdrawal or prevention" of a reward (the definition). Therefore, when annotating, if they've felt that the research would benefit or help further understand Frustrative Non-Reward and its associated behaviors, they've annotated it as related (this included environmental, social, and biological factors influencing relational and physical aggression).

Over thirty teams registered to participate in at least one of the RDoC tasks. Eventually, 5 teams submitted their predictions; four teams submitted for both tasks and one team for only task 1. In the following analysis, we will be using the unique team identifiers (assigned during the task registration[16]) for referring to the 5 teams. Note that these team identifiers bear no significance other than identifying different teams.

3.1 Task 1: Information Retrieval

Four teams submitted their predictions for this task and their scores are reported in Table 7. Bold entries indicate the highest score for the corresponding construct. Although included in Table 7, we excluded the two constructs, *Circadian Rhythms* and *Sleep and Wakefulness*, from the final analysis since these constructs contain one and zero negative articles, respectively, leading to perfect performance (see Table 5). Team 30 achieved the highest mean average precision (0.86) among all teams. Though Team 10 achieved the second-highest mean average precision (0.85) that is very close to the highest, we found a statistically significant difference between the scores of these two teams (paired t-test, p=0.005, $\alpha = 0.05$). Team 30 achieved the highest scores for *Frustrative Nonreward*, *Loss* and *Potential Threat (Anxiety)* whereas Team 10 achieved the highest scores for the other three constructs. Though it seems the

[16] https://sites.google.com/view/rdoc-task/registration

scores achieved by the Team 10 and 30 is close to the baseline, we found these scores to be statistically significantly higher from the baseline for both Team 10 (paired t-test, p=0.022) and Team 30 (paired t-test, p=0.043) using $\alpha = 0.05$.

The last column in Table 7 reports the average score for the corresponding construct. It is seemingly easier to rank the relevant articles for *Arousal* and *Potential Threat (Anxiety)* whereas it is moderately difficult for *Sustained Threat*. Sustained Threat being more challenging for IR may be explained by the fact that the annotators also found it to be the most challenging construct for task 1 annotation.

3.2 Task 2: Sentence Extraction

Five teams submitted their predictions for this task and their scores are reported in Table 8. Bold entries indicate the highest score for the corresponding construct. Team 30 again achieved the highest macro average accuracy (0.58) among all the teams and the highest score for five out of eight constructs. Team 7 achieved the highest score for the rest of the three constructs with significant improvement over Team 30. Construct-wise highest scores of *Sustained Threat*, *Arousal* and *Circadian Rhythms*, achieved by either Team 7 or Team 30, are higher by about 0.27 compared to the baseline performance. In addition, the highest scores for other constructs are also higher by more than 0.17 compared to the baseline performance.

Frustrative Nonreward has the lowest average score (0.31) among all the constructs. Moreover, its highest score (0.43) is also the lowest among all the highest scores. So, extracting the most relevant sentences for *Frustrative Nonreward* is seemingly more difficult compared to the other constructs.

Typically, participating teams performed relatively better on shorter abstracts (see Table 9), which is intuitive due to that fact the models have a higher chance of finding the most similar sentences for shorter abstracts. Similarly, they performed well for abstracts with more gold-standard sentences (see Table 10). This is also intuitive because when there are more gold-standard sentences, there is a higher chance of matching one of them.

4 Conclusion and Future work

We introduced a novel mental health informatics task called RDoC task at this years BioNLP-OST 2019 workshop. RDoC task is a combination of two subtasks on information retrieval and sentence extraction using the RDoC framework. Originally, over 30 teams registered, highlighting a significant interested in mental health informatics and/ or RDoC. Eventually, four and five teams participated in the information retrieval and sentence extraction tasks, respectively.

Overall results show that the top-performing team was able to easily outperform the baseline models for most of the constructs. On the other hand, the baseline methods outperform at least one system (often more). This is surprising given that the baseline models are not sophisticated. One reason could be that the baseline methods do not utilize training data, while the participating methods may have been overfitted to the training data. Another reason could be, these simple baselines perform better than (most likely more complex) participating models due to working with shorter documents (i.e. abstracts). If the full texts were made available, models primarily depended on TFIDF may struggle to achieve good performance. Regardless, this calls for more sophisticated methods for both tasks because any other sophisticated method (such as Lucene [17] or MetaMap [2]) used a baseline may have outperformed even more participating teams.

The publicly made available gold-standard data should serve as a valuable resource for the brain research/ mental health and RDoC researchers and curators going forward. In the future iterations of the RDoC task, we would like to incorporate either all available or a well-representative set of RDoC constructs covering all domains. We plan to improve the quality of benchmark data using "reconciliation" instead of "majority voting" as well as using improved search that uses MeSH and/ or other vocabularies.

And equally important aspect would be to explore information extraction tasks such as extracting various entities under different RDoC units of analysis, which is likely more useful for the curators. This would also mean an exploration of incorporating full text in addition to abstracts will be required due to the abundance of entities existing in the full articles compared to just the abstract. Last but not least, exploring clever ways to maintain the enthusiasm of the registered teams would be highly valuable to the overall success of the future iterations of the RDoC task .

Table 7: Performance of retrieving PubMed Abstracts related to the corresponding RDoC construct (Task 1). Four teams participated (T10, T21, T22, and T30). IQR: inter-quartile range. Bolded scores are the highest across all teams per the construct.

RDoC construct	Baseline	T10	T21	T22	T30	Avg	IQR
Acute Threat (Fear)	0.74	**0.89**	0.83	0.67	0.85	0.81	0.17
Potential Threat (Anxiety)	0.90	0.87	0.89	0.81	**0.94**	0.88	0.10
Frustrative Nonreward	0.70	0.69	0.67	0.61	**0.73**	0.68	0.10
Sustained Threat	0.64	**0.64**	**0.64**	0.41	0.63	0.58	0.18
Loss	0.77	0.74	0.71	0.61	**0.78**	0.71	0.14
Arousal	0.95	**0.93**	0.91	0.84	0.92	0.90	0.07
Circadian Rhythms	1.00	1.00	1.00	1.00	1.00	1.00	0.00
Sleep and Wakefulness	1.00	1.00	1.00	0.98	1.00	1.00	0.02
Mean Average Precision	0.84	0.85	0.83	0.74	**0.86**	–	–

Table 8: Performance of extracting the most relevant sentence from each abstract related to the corresponding RDoC construct (Task 2). Five teams participated (T7, T10, T21, T22, and T30). IQR: inter-quartile range.

RDoC construct	Baseline	T7	T10	T21	T22	T30	Avg	IQR
Acute Threat (Fear)	0.53	0.58	0.68	0.37	0.47	**0.74**	0.57	0.29
Potential Threat (Anxiety)	0.41	0.41	0.32	0.15	0.38	**0.59**	0.37	0.27
Frustrative Nonreward	0.23	**0.43**	0.34	0.11	0.29	0.37	0.31	0.20
Sustained Threat	0.19	**0.47**	0.36	0.14	**0.47**	0.42	0.37	0.22
Loss	0.53	0.26	0.56	0.26	0.62	**0.74**	0.49	0.42
Arousal	0.46	0.46	0.62	0.12	0.42	**0.73**	0.47	0.41
Circadian Rhythms	0.43	**0.70**	0.47	0.10	0.60	0.47	0.47	0.37
Sleep and Wakefulness	0.43	0.33	0.50	0.17	0.57	**0.60**	0.43	0.34
Macro Average Accuracy	0.40	0.46	0.48	0.18	0.48	**0.58**	–	–

Table 9: Variation of Accuracy over various size of abstract. #*m*-*n*: abstracts with *m* to *n* sentences.

RDoC construct	#3-8	#9-14	#15-20
Acute Threat	0.60	0.64	0.40
Potential Threat	0.47	0.39	–
Frustrative Nonreward	0.28	0.25	0.50
Sustained Threat	0.39	0.32	0.40
Loss	0.62	0.60	0.31
Arousal	0.53	0.39	–
Circadian Rhythms	0.38	0.54	0.00
Sleep & Wakefulness	0.58	0.42	–

Table 10: Variation of Accuracy over the number of most relevant (gold-standard) sentences in abstracts. #x: abstracts with x relevant (gold-standard) sentences.

RDoC construct	#1	#2	#3	#4
Acute Threat	0.29	0.60	0.69	0.80
Potential Threat	0.31	0.47	0.57	0.75
Frustrative Nonreward	0.18	0.35	0.54	–
Sustained Threat	0.29	0.42	0.37	0.25
Loss	0.56	0.65	0.50	–
Arousal	0.47	0.45	0.65	0.40
Circadian Rhythms	0.17	0.41	0.60	0.61
Sleep & Wakefulness	0.23	0.56	0.67	0.80

Acknowledgments

This work was partially funded by The Center for Mental Health Research and Recovery (CMHRR) at Montana State University (MSU). We would like to thank Robell Basset, Lenin Lewis, Ninoo De Silva, and Hannah Reiser (from the Department of Psychology, MSU), and Soumilee Chaudhuri (from the Department of Cell Biology & Neuroscience, MSU) for assisting the curation process.

References

[1] Dean Carcone and Anthony C Ruocco. Six years of research on the National Institute of Mental Healths Research Domain Criteria (RDoC) initiative: a systematic review. *Frontiers in cellular neuroscience*, 11:46, 2017.

[2] K Bretonnel Cohen, Tom Christiansen, and Lawrence E Hunter. Metamap is a superior baseline to a standard document retrieval engine for the task of finding patient cohorts in clinical free text. In *TREC*. Citeseer, 2011.

[3] Kevin Bretonnel Cohen and Dina Demner-Fushman. *Biomedical natural language processing*, volume 11. John Benjamins Publishing Company, 2014.

[4] Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, 2015.

[5] Bruce N Cuthbert. The rdoc framework: facilitating transition from icd/dsm to dimensional approaches that integrate neuroscience and psychopathology. *World Psychiatry*, 13(1):28–35, 2014.

[6] Bruce N Cuthbert and Thomas R Insel. Toward the future of psychiatric diagnosis: the seven pillars of rdoc. *BMC medicine*, 11(1):126, 2013.

[7] Hong-Jie Dai, Emily Chia-Yu Su, Mohy Uddin, Jitendra Jonnagaddala, Chi-Shin Wu, and Shabbir Syed-Abdul. Exploring associations of clinical and social parameters with violent behaviors among psychiatric patients. *Journal of biomedical informatics*, 75:S149–S159, 2017.

[8] Daniel Dopp, Adam Morrone, and Indika Kahanda. KinDER: A biocuration tool for extracting kinase knowledge from biomedical literature. *Proceedings of the BioCreative VI Workshop*, Oct 2017.

[9] Michele Filannino, Amber Stubbs, and Özlem Uzuner. Symptom severity prediction from neuropsychiatric clinical records: Overview of 2016 cegs n-grid shared tasks track 2. *Journal of biomedical informatics*, 75:S62–S70, 2017.

[10] International Society for Biocuration. Biocuration: Distilling data into knowledge. *PLOS Biology*, 16(4):1–8, 04 2018.

[11] Julien Gobeill, Pascale Gaudet, Daniel Dopp, Adam Morrone, Indika Kahanda, Yi-Yu Hsu, Chih-Hsuan Wei, Zhiyong Lu, and Patrick Ruch. Overview of the biocreative vi text-mining services for kinome curation track. *Database*, 2018(1):bay104, 2018.

[12] Chung-Chi Huang and Zhiyong Lu. Community challenges in biomedical text mining over 10 years: success, failure and the future. *Briefings in bioinformatics*, 17(1):132–144, 2015.

[13] Jessica I Lake, Cindy M Yee, and Gregory A Miller. Misunderstanding rdoc. *Zeitschrift für Psychologie*, 2017.

[14] Scott O Lilienfeld and Michael T Treadway. Clashing diagnostic approaches: Dsm-icd versus rdoc. *Annual review of clinical psychology*, 12:435–463, 2016.

[15] Inderjeet Mani. *Advances in automatic text summarization*. MIT press, 1999.

[16] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Evaluation in information retrieval*, page 139161. Cambridge University Press, 2008.

[17] Michael McCandless, Erik Hatcher, and Otis Gospodnetic. *Lucene in action: covers Apache Lucene 3.0*. Manning Publications Co., 2010.

[18] David N Milne, Glen Pink, Ben Hachey, and Rafael A Calvo. Clpsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 118–127, 2016.

[19] Malvina Nissim, Lasha Abzianidze, Kilian Evang, Rob van der Goot, Hessel Haagsma, Barbara Plank, and Martijn Wieling. Sharing is caring: The future of shared tasks. *Computational Linguistics*, 43(4):897–904, 2017.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[21] John P Pestian, Pawel Matykiewicz, and Michelle Linn-Gust. What's in a note: construction of a suicide note corpus. *Biomedical informatics insights*, 5:BII–S10213, 2012.

[22] John P Pestian, Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, Jan Wiebe, K Bretonnel Cohen, John Hurdle, and Christopher Brew. Sentiment analysis of suicide notes: A shared task. *Biomedical informatics insights*, 5:BII–S9042, 2012.

[23] Anand Rajaraman and Jeffrey David Ullman. *Data Mining*, page 117. Cambridge University Press, 2011.

[24] Justus J Randolph. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*, 2005.

[25] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, April 2009.

[26] Eric W. Sayers, Tanya Barrett, Dennis A. Benson, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 38(suppl_1):D5–D16, 1 2010.

[27] Amit Singhal et al. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.

[28] Amber Stubbs, Michele Filannino, and Özlem Uzuner. De-identification of psychiatric intake records: Overview of 2016 cegs n-grid shared tasks track 1. *Journal of biomedical informatics*, 75:S4–S18, 2017.

[29] Tung Tran and Ramakanth Kavuluru. Predicting mental conditions based on history of present illness in psychiatric notes with deep neural networks. *Journal of biomedical informatics*, 75:S138–S148, 2017.

[30] Özlem Uzuner, Amber Stubbs, and Michele Filannino. A natural language processing challenge for clinical records: Research domains criteria (RDoC) for psychiatry. *Journal of biomedical informatics*, 75:S1–S3, 2017.

[31] Ellen M Voorhees, Donna K Harman, et al. *TREC: Experiment and evaluation in information retrieval*, volume 63. MIT press Cambridge, 2005.

[32] Yaoyun Zhang, Olivia Zhang, Yonghui Wu, Hee-Jin Lee, Jun Xu, Hua Xu, and Kirk Roberts. Psychiatric symptom recognition without labeled data using distributional representations of phrases and on-line knowledge. *Journal of biomedical informatics*, 75:S129–S137, 2017.

APPENDIX B

AUTOMATICALLY CATALOGING SCHOLARLY ARTICLES USING LIBRARY OF CONGRESS SUBJECT HEADINGS

Automatically Cataloging Scholarly Articles using Library of Congress Subject Headings

Nazmul Kazi[1], **Nathaniel Lane**[1], **Indika Kahanda**[2]
[1] Montana State University, MT, USA
[2] University of North Florida, FL, USA
kazinazmul.hasan@montana.edu
nathaniel.lane@student.montana.edu
indika.kahanda@unf.edu

Abstract

Institutes are required to catalog their articles with proper subject headings so that the users can easily retrieve relevant articles from the institutional repositories. However, due to the rate of proliferation of the number of articles in these repositories, it is becoming a challenge to manually catalog the newly added articles at the same pace. To address this challenge, we explore the feasibility of automatically annotating articles with Library of Congress Subject Headings (LCSH). We first use web scraping to extract keywords for a collection of articles from the Repository Analytics and Metrics Portal (RAMP). Then, we map these keywords to LCSH names for developing a gold-standard dataset. As a case study, using the subset of Biology-related LCSH concepts, we develop predictive models by formulating this task as a multi-label classification problem. Our experimental results demonstrate the viability of this approach for predicting LCSH for scholarly articles.

1 Introduction

An Institutional Repository (IR) is the collection of scholarly work hosted and maintained by institutions such as universities. For example, "ScholarWorks[1] is an open access repository for the capture of the intellectual work of Montana State University (MSU) in support of its teaching and research goals". Repository Analytics and Metrics Portal (RAMP) is a web service that accurately counts item downloads for each article in the institutional repository (Obrien et al., 2016; OBrien et al., 2017). Besides counting the number of downloads, RAMP stores metadata of the articles such as title, abstract, and keywords. Currently, nearly 40 institutions have registered their repositories with RAMP.

[1]https://scholarworks.montana.edu/

To facilitate the easy finding of articles, the IR managers need to catalog them using different subject headings manually. One of the most popular vocabularies for cataloging is the Library of Congress Subject Headings (LCSH) (Walsh, 2011). LCSH is a subject indexing language that is actively maintained since 1898 to catalog materials in the Library of Congress and most widely adopted by large and small libraries around the world (Work, 2016). A subject heading is the most specific word or a group of words that capture the essence of a subject category. Due to the rapid growth of items in IRs, manual cataloging using LCSH or other vocabularies is becoming highly resource-consuming (Engelson, 2013).

Due to the above challenge, there have been a few previous attempts on the automatic assignment of LCSH through keyword extraction (Wartena et al., 2010; Aga et al., 2016), by collecting LCSH concepts that are assigned to similar texts (Paynter, 2005), using semantic similarity (Yi, 2010), and co-occurrence-based mapping (Vizine-Goetz et al., 2004). These techniques primarily depend on the presence of the keywords or similar words/ phrases within the actual text and do not utilize machine learning. Furthermore, one of the studies claims that the prediction of LCSH using machine learning may be infeasible due to the large size of the vocabulary leading to inadequate training data (Wartena et al., 2010). Note that machine learning has been used for a seemingly similar but actually different task of predicting Library of Congress Classification (LCC) (Frank and Paynter, 2004). However, despite the similarity in their names, LCC and LCSH are completely different vocabularies.

Semantic indexing with other vocabularies has gained traction recently (Mirowski et al., 2010; Salakhutdinov and Hinton, 2009; Wu et al., 2014). Most notably, predicting Medical Subject Headings (MeSH) for biomedical literature using machine

Proceedings of the 16th Conference of the European Chapter of the Associationfor Computational Linguistics: Student Research Workshop, pages 43–49
April 19 - 23, 2021.

learning and deep learning techniques has seen significant recent interest (Mao and Lu, 2017; Jin et al., 2018; Kehoe et al., 2017; Rios and Kavuluru, 2015; Kosmopoulos et al., 2015; Yan et al., 2016) thanks to the BioASQ challenge on Biomedical Semantic Indexing (Tsatsaronis et al., 2015).

In this work, we explore the feasibility of developing an automated pipeline for predicting LCSH for scholarly articles using machine learning. As a case study, we leverage an extensive collection of scholarly articles from RAMP and generate a gold-standard dataset by assigning Biology-related LCSH concepts to each article through web scraping and string matching techniques. Using this gold-standard data, we develop predictive models that can predict LCSH by modeling this as a multi-label classification problem. Our experimental results indicate the effectiveness of the proposed approach.

2 Methodology

2.1 Data

In this approach, we build a gold-standard dataset by scraping RAMP data from 27 institutional repositories (IRs). A high-level overview of our approach is shown in Figure 1.

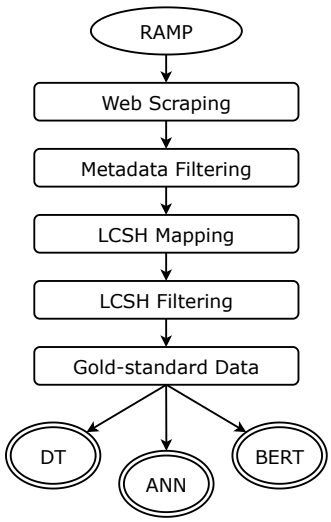

Figure 1: A high-level overview of our approach.

We identify the citable content downloads (CCD) from each institutional repository (IR) between July 2017 and July 2018 . Then, we scrape all metadata of each CCD from RAMP for the subset that includes all unique CCDs.

The raw data (scraped from RAMP) contains 457,879 articles and 270 different metadata types. However, we use only *title* concatenated with *abstract*, *article type*, and *keywords* for this study, and discard other metadata. There are many reasons why some of the metadata are empty. For example, items such as newspapers do not include abstracts, and sometimes IR managers add items into repositories without populating metadata. Therefore, we first discard articles without a title, an abstract, or keywords, which reduces the dataset to 126,655 articles that have a title, an abstract, and at least one keyword. Then, we map each keyword to the subject names from the 41st edition of LCSH[2] using full string matching (case insensitive). If a keyword does not match with any subject, we ignore that keyword.

Any article without at least one assigned subject heading is discarded. This results in a smaller set of articles with annotated subject headings. Then, we filter out any subjects not related to Biology by only retaining the concept *Biology* (sh85014203)[3] and its descendants. Finally, we remove subject headings that are annotated to less than 100 articles. After all the above, we have a dataset composed of 17,367 articles with 66 Biology-related subject headings. This LCSH-annotated dataset is used as the gold-standard dataset for developing predictive models. Note that while the string matching technique used in this study itself can potentially be used for "predicting" LCSH terms, we are assuming that unseen items that need to be annotated with LCSH in real-life may not necessarily come with keywords (and hence we resort to developing predictive machine learning models). The distribution of articles across IRs in this dataset is shown in Table 1.

2.2 Models

We model the task of predicting LCSH concepts as a multi-label classification problem and develop three supervised machine learning models using the above generated gold-standard data. These models are 1) Decision Tree (DT), 2) Artificial Neural Networks (ANN), and 3) Bidirectional Encoder Representations from Transformers (BERT). All the models are implemented using scikit-learn[4], Ten-

[2]https://loc.gov/aba/publications/FreeLCSH/freelcsh.html
[3]http://id.loc.gov/authorities/subjects/sh85014203.html
[4]https://scikit-learn.org/

	IR Name	# Articles
1	Deep Blue	7,820
2	DRUM	1,578
3	EASP	1,171
4	UWSpace	1,063
5	OpenBU	960
6	MacSphere	917
7	Texas ScholarWorks	849
8	Mountain Scholar	631
9	Epsilon Open Archive	576
10	K-REx	464
11	MSU ScholarWorks	405
12	OAKTrust	380
13	MD-SOAR	245
14	SHAREOK	192
15	Others	116
	Total:	17,367

Table 1: Number of articles per institute in the gold-standard dataset.

sorFlow[5], Transformers[6] and PyTorch[7] libraries. In our preliminary work, We also train models using Support Vector Machines and Random Forest classifiers, but none of them perform better than the models reported in this paper (data not shown).

We choose standard but varying pre-processing steps independently for each model since certain pre-processing techniques work well for some models over the others. For example, removing stopwords is a common practice for Decision Tree models but not for BERT since stopwords typically can act as noise for the former.

2.2.1 Decision Tree (DT) model

We apply the Decision Tree classifier to develop a tree-based one-vs-rest classification model. We use TF-IDF (term frequency-inverse document frequency) vectorizer with a word-based analyzer for feature extraction. We use lemmatization and stop word removal as standard pre-processing steps. We include both uni-grams and bi-grams as features and train our model over the top 10,000 features. Our model returns a binary value, i.e., either 0 or 1, as the prediction.

2.2.2 Artificial Neural Network (ANN) model

For the shallow artificial neural network model, we use the TF-IDF scores as input. These are generated using scikit-learn's TfidfVectorizer class. All stop words (common words such as "the" or "and") are removed before vectorization, and only the terms that appear in a minimum of 1% of all documents are kept.

Our artificial neural network has four layers: an input layer with 2,251 nodes, a dropout layer with a rate of 0.1, a hidden layer with 132 nodes, and an output layer with 66 nodes (one for each label) with a sigmoid activation function. We initially experimented with many different network structures but ultimately find that a single hidden layer with 132 nodes, double the number in the output, produces the best results (data not shown). We use 5-fold nested cross-validation to find the optimal epoch for training the networks. We train the largest network with 100 epochs and find 10 epochs as optimal as the learning curve reaches convergence. We use this optimal epoch to train all networks.

2.2.3 Bidirectional Encoder Representations from Transformers (BERT) model

We use the pre-trained BERT-Base (uncased) model (Devlin et al., 2018) and fine-tune it for multi-label text classification. The base model has 12 transformer blocks, i.e., hidden layers, a hidden size of 768, 12 attention heads, and 110 million parameters (Devlin et al., 2018). The model is pre-trained for English on uncased Wikipedia and BooksCorpus. For fine-tuning the model, we use *Adam* optimizer with a learning rate of $2e-5$, $\epsilon = 1e-8$, L2 weight decay of 0.01, learning rate warmup over the first 500 steps with linear decay and Cross-Entropy Loss function. We observe the learning curve over 5-fold nested cross-validation and find 6 epochs as the optimal number. Any example longer than the 512 token length restriction enforced by the BERT-Base model is truncated.

2.3 Experimental Setup and Metrics

In order to obtain unbiased estimations of model performance, we evaluate our models using 5-times 5-fold stratified cross-validation (Sechidis et al., 2011; Szymański and Kajdanowicz, 2017). We primarily report the performances of our models using Maximum F1-score (F_{max}), Precision at F_{max} and Recall at F_{max}. Precision reports the percentage of true samples among the samples that have been predicted as true, whereas Recall reports the percentage of true samples retrieved by the model. F1-score is the harmonic mean of precision and re-

[5]https://www.tensorflow.org/
[6]https://huggingface.co/transformers/
[7]https://pytorch.org/

Subject Frequency	# subjects	DT			ANN			BERT		
		P	R	F1	P	R	F_{max}	P	R	F_{max}
$[100, 200)$	35	0.36	0.35	0.36	0.48	0.40	0.43	0.51	0.43	0.43
$[200, 300)$	15	0.40	0.39	0.39	0.48	0.46	0.47	0.56	0.51	0.49
$[300, 400)$	6	0.31	0.30	0.30	0.42	0.44	0.43	0.55	0.55	0.54
$[400, 900)$	7	0.41	0.41	0.41	0.48	0.57	0.52	0.59	0.70	0.64
$[1700, 2600]$	3	0.40	0.40	0.40	0.46	0.67	0.54	0.57	0.71	0.63
Macro average:		0.38	0.37	0.37	0.46	0.51	0.48	0.56	0.58	0.55

Table 2: Model performance per subject frequency range. # subjects: Number of unique subjects within the range, P: precision, R: recall.

Article Type	Freq.	Length		Average Number of		F_{max}		
		Avg	Std	Keywords	Subjects	DT	ANN	BERT
Thesis	6,765	379.89	191.32	30.24	1.15	0.31	0.38	0.41
Article	1,077	225.80	99.52	21.25	1.29	0.19	0.23	0.24
Report	880	442.89	280.80	15.80	1.09	0.14	0.18	0.19
Paper	364	207.68	111.74	22.29	1.17	0.10	0.12	0.16
Book	48	221.31	194.41	20.27	1.08	0.02	0.03	0.04
Others	383	164.54	116.59	24.53	1.30	0.12	0.14	0.19
NA	7,850	253.99	147.47	21.52	1.27	0.30	0.38	0.41

Table 3: Model performance per article type. NA: Not Available, Freq: number of articles in type, Length: number of words in title and abstract, P: precision, and R: recall.

Subject	Freq	F_{max}		
		DT	ANN	BERT
Commencement ceremonies	141	0.99	1.00	1.00
Discrimination	227	0.83	0.88	0.88
Irrigation	125	0.72	0.66	0.89
Machine Learning	260	0.68	0.71	0.75
Nanoparticles	174	0.67	0.67	0.78
Self-efficacy	112	0.64	0.69	0.71
Animal ecology	520	0.56	0.67	0.79
Autism	103	0.68	0.51	0.75
Feminism	113	0.63	0.52	0.76
Planning	245	0.50	0.65	0.69

Table 4: Top ten easiest to predict subjects. Freq: Frequency of subject in the dataset.

Subject	Freq	F_{max}		
		DT	ANN	BERT
Social psychology	157	0.05	0.14	0.02
Clinical psychology	196	0.10	0.22	0.00
Metabolism	104	0.14	0.19	0.00
Molecular biology	185	0.07	0.15	0.11
Developmental psychology	174	0.14	0.20	0.00
Cognition	109	0.18	0.20	0.00
Epidemiology	224	0.17	0.20	0.04
Zoology	242	0.13	0.24	0.05
Physiology	190	0.12	0.20	0.11
Neurology	176	0.23	0.25	0.00

Table 5: Top ten hardest to predict subjects. Freq: Frequency of subject in the dataset.

call. Unlike F1, F_{max}, which is computed across a range of thresholds, is threshold independent. More specifically, let threshold $t \in [0, 1]$, then

$$F_{max} = \max_t \left\{ \frac{2 \cdot \text{precision}(t) \cdot \text{recall}(t)}{\text{precision}(t) + \text{recall}(t)} \right\}$$

For this study, we use a step size of 0.05 for thresholds and Macro-averaging (arithmetic mean) for aggregating the performance across classes.Note that since the DT model returns binary predictions directly, without class probabilities, we report the performance of this model only using F1 instead of F_{max}.

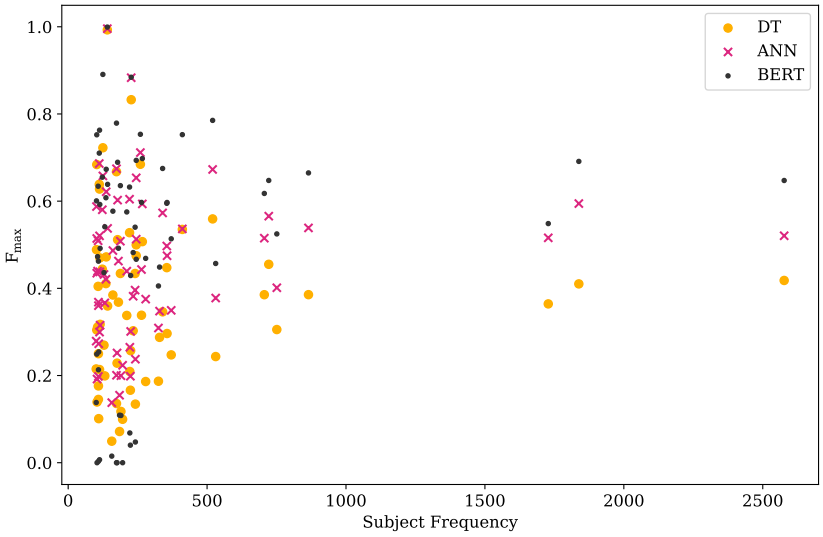

Figure 2: Model performance against subject frequency. DT: Decision Tree, ANN: Artificial Neural Network.

3 Results and Discussion

The overall performance for all our models is depicted in Table 2. Overall, the BERT model performs the best, and the DT model performs the worst among the three models. The DT model achieves an average F1 score of 0.37, whereas the lowest F1 score (0.30) is observed for frequency range $[300, 400)$. The performance of the DT model is seemingly immune to the frequency of subjects. The ANN model notably outperforms the DT model with an average F_{max} of 0.48. The ANN model also struggles for frequency range $[300, 400)$. However, the lowest F_{max} (0.43) of ANN is higher than the best F1 score (0.40) achieved by DT in any frequency range. Except for frequency range $[300, 400)$, we can see an increase in F_{max} of ANN as the frequency range increases. The BERT model significantly outperforms both DT and ANN models with an average F_{max} of 0.55 and shows a positive correlation between F_{max} and frequency range.

Figure 2 shows variation of performance of all three against the frequency. The subjects between range $[100, 200)$ are widely spread across the y-axis (F_{max}) for each model, which indicates that the easiest and the hardest subject to predict have similar subject frequencies. Top ten easiest and hardest subjects across all three models are listed in Table 4 and Table 5, respectively. We use macro-averaged F-score from all three models to compile these rankings. All three models show their best performance for the same subject, *Commencement ceremonies*. Both DT and ANN have a non-zero F-score for each subject. Despite being the best model, BERT shows zero F_{max} for several subjects, e.g., *Clinical psychology*.

We also assess the performance of each model per document type, as reported in Table 3. For the following analysis, we exclude the document type denoted as NA for which the corresponding metadata was missing. Same as before, BERT performs the best, and ANN outperforms DT. All three models show their best and worst performance for the same article types across all models, Thesis and Book, respectively. The frequency of each type may have played a significant role in these extremes. This is further supported by the fact that the performance across all three models follows the same trend: as the frequency decreases, the performance decreases as well.

4 Conclusions and Future Work

In this work, we explore the feasibility of using machine learning for predicting LCSH for scholarly articles. We first generate a gold-standard dataset annotated with LCSH subjects by web scraping/ string matching and utilize this data for developing multi-label classification models. Our results indicate the feasibility of our approach. We believe our approach is applicable to other data similar to LCSH concepts. This automated pipeline should

be extremely valuable to librarians for expediting the manual cataloging process. We plan to measure the efficiency gains of this method through the Montana State University Library.

While our approach displays promising results, there are many different avenues for future investigation. First, in this work, we map the web scraped keywords to subject names (instead of identifiers or IDs). However, some subject names may map to more than one identifier (e.g., Psychology: sh85108459 or sh2002011487). So, we plan to explore two different solutions to this. One approach is to develop a chain-classifier that can predict the LCSH IDs using the already predicted subjects (i.e., a second classifier for disambiguation). Another option is the improve the web scraping/ string matching pipeline so that we can generate a gold-standard dataset directly annotated with IDs.

To improve the performance of our traditional machine learning models, we plan to investigate the inclusion of hand-engineered features, other resources such as MeSH terms, metadata fields that were ignored in this study, and the hierarchical information from the LCSH. Besides, using larger more sophisticated language models (e.g., Megatron-LM), using the complete set of LCSH terms (without restricting to Biology-related), and structured output models that explicitly use the hierarchy information will likely improve performance. Moreover, Extreme Multi-Label (XML) models that are equipped to handle very large sets of classes (Kumar et al., 2019) will also likely provide better performance.

5 Acknowledgement

We would like to thank Patrick OBrien and Kenning Arlitsch from Montana State University Library and Jonathan Wheeler from University of New Mexico for providing us with the data and guidance for this project. This work was supported in part by NSF awards CNS-1730158, ACI-1540112, ACI-1541349, OAC-1826967, the University of California Office of the President, and the University of California San Diego's California Institute for Telecommunications and Information Technology/Qualcomm Institute. Thanks to CENIC for the 100Gpbs networks.

References

Rosa Tsegaye Aga, Christian Wartena, and Michael Franke-Maier. 2016. Automatic recognition and disambiguation of library of congress subject headings. In *International Conference on Theory and Practice of Digital Libraries*, pages 442–446. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Leslie Engelson. 2013. Correlations between title keywords and lcsh terms and their implication for fast-track cataloging. *Cataloging & classification quarterly*, 51(6):697–727.

Eibe Frank and Gordon W Paynter. 2004. Predicting library of congress classifications from library of congress subject headings. *Journal of the American Society for Information Science and Technology*, 55(3):214–227.

Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu. 2018. Attentionmesh: Simple, effective and interpretable automatic mesh indexer. In *Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*, pages 47–56.

Adam K Kehoe, Vetle I Torvik, Matthew B Ross, and Neil R Smalheiser. 2017. Predicting mesh beyond medline. In *Proceedings of the 1st Workshop on Scholarly Web Mining*, pages 49–56.

Aris Kosmopoulos, Ion Androutsopoulos, and Georgios Paliouras. 2015. Biomedical semantic indexing using dense word vectors in bioasq. *J BioMed Semant Suppl BioMedl Inf Retr*, 3410:959136040–1510456246.

P. Kumar, V. K. Dubey, and M. I. H. Showrov. 2019. A comparative analysis on various extreme multi-label classification algorithms. In *2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)*, pages 265–268.

Yuqing Mao and Zhiyong Lu. 2017. Mesh now: automatic mesh indexing at pubmed scale via learning to rank. *Journal of biomedical semantics*, 8(1):15.

Piotr Mirowski, M Ranzato, and Yann LeCun. 2010. Dynamic auto-encoders for semantic indexing. In *Proceedings of the NIPS 2010 Workshop on Deep Learning*, volume 2.

Patrick OBrien, Kenning Arlitsch, Jeff Mixter, Jonathan Wheeler, and Leila Belle Sterman. 2017. Ramp–the repository analytics and metrics portal. *Library Hi Tech*.

Patrick Obrien, Kenning Arlitsch, Leila Sterman, Jeff Mixter, Jonathan Wheeler, and Susan Borda. 2016. Undercounting file downloads from institutional repositories. *Journal of Library Administration*, 56(7):854–874.

Gordon W Paynter. 2005. Developing practical automatic metadata assignment and evaluation tools for internet resources. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'05)*, pages 291–300. IEEE.

Anthony Rios and Ramakanth Kavuluru. 2015. Analyzing the moving parts of a large-scale multi-label text classification pipeline: Experiences in indexing biomedical articles. In *2015 International Conference on Healthcare Informatics*, pages 1–7. IEEE.

Ruslan Salakhutdinov and Geoffrey Hinton. 2009. Semantic hashing. *International Journal of Approximate Reasoning*, 50(7):969–978.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. *Machine Learning and Knowledge Discovery in Databases*, pages 145–158.

Piotr Szymański and Tomasz Kajdanowicz. 2017. A network perspective on stratification of multi-label data. In *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, volume 74 of *Proceedings of Machine Learning Research*, pages 22–35, ECML-PKDD, Skopje, Macedonia. PMLR.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138.

Diane Vizine-Goetz, Carol Hickey, Andrew Houghton, and Roger Thompson. 2004. Vocabulary mapping for terminology services. *Journal of digital information*, 4(4):2004.

John Walsh. 2011. The use of library of congress subject headings in digital collections. *Library review*.

Christian Wartena, Rogier Brussee, and Wout Slakhorst. 2010. Keyword extraction using word co-occurrence. In *2010 Workshops on Database and Expert Systems Applications*, pages 54–58. IEEE.

Jill A Work. 2016. Legislating librarianship. *The Political Librarian*, 2(2):7.

Hao Wu, Martin Renqiang Min, and Bing Bai. 2014. Deep semantic embedding. In *SMIR@ SIGIR*.

Yan Yan, Xu-Cheng Yin, Bo-Wen Zhang, Chun Yang, and Hong-Wei Hao. 2016. Semantic indexing with deep learning: a case study. *Big Data Analytics*, 1(1):1–13.

Kwan Yi. 2010. A semantic similarity approach to predicting library of congress subject headings for social tags. *Journal of the American Society for Information Science and Technology*, 61(8):1658–1672.

APPENDIX C

PSYCHIATRY TRANSCRIPT ANNOTATION:

PROCESS STUDY AND IMPROVEMENTS

Psychiatry Transcript Annotation: Process Study and Improvements

Srinivasan Sridhar, MS[1]; Nazmul Kazi[2]; Indika Kahanda, PhD[3];
Bernadette McCrory, PhD, MPH, PE, CHFP[1]

[1]Mechanical and Industrial Engineering, Montana State University, Bozeman, MT
[2]Gianforte School of Computing, Montana State University, Bozeman, MT
[3]School of Computing, University of North Florida, Jacksonville, FL

Background: The demand for psychiatry is increasing each year. Limited research has been performed to improve psychiatrist work experience and reduce daily workload using computational methods. There is currently no validated tool or procedure for the mental health transcript annotation process for generating "gold-standard" data. The purpose of this paper was to determine the annotation process for mental health transcripts and how it can be improved to acquire more reliable results considering human factors elements. **Method:** Three expert clinicians were recruited in this study to evaluate the transcripts. The clinicians were asked to fully annotate two transcripts. An additional five subjects were recruited randomly (aged between 20-40) for this pilot study, which was divided into two phases, phase 1 (annotation without training) and phase 2 (annotation with training) of five transcripts. Kappa statistics were used to measure the inter-rater reliability and accuracy between subjects. **Results:** The inter-rater reliability between expert clinicians for two transcripts were 0.26 (CI 0.19 to 0.33) and 0.49 (CI 0.42 to 0.57), respectively. In the pilot testing phases, the mean inter-rater reliability between subjects was higher in phase 2 with training transcript (k= 0.35 (CI 0.052 to 0.625)) than in phase 1 without training transcript (k= 0.29 (CI 0.128 to 0.451)). After training, the accuracy percentage among subjects was significantly higher in transcript A (p=0.04) than transcript B (p=0.10). **Conclusion:** This study focused on understanding the annotation process for mental health transcripts, which will be applied in training machine learning models. Through this exploratory study, the research found appropriate categorical labels that should be included for transcripts annotation, and the importance of training the subjects. Contributions of this case study will help the psychiatric clinicians and researchers in implementing the recommended data collection process to develop a more accurate artificial intelligence model for fully- or semi-automated transcript annotation.

INTRODUCTION

The demand for psychiatric treatment is increasing each year in the United States (US) (Butryn, Bryant, Marchionni, & Sholevar, 2017). With an increase in the number of mental illness cases, the shortage of psychiatrists is becoming a major issue in the US (Butryn et al., 2017). According to the Board of American Medical Colleges, it is estimated that in the US, by 2024, there will be only 11.3 psychiatrists per 100,000 people (Satiani, Niedermier, Satiani, & Svendsen, 2018). The increasing demand for psychiatrists will directly increase psychiatrists' workload and daily work hours (McCluskey, 2019). Prior researchers have explored the concepts of Artificial Intelligence (AI) and Natural Language Processing (NLP) to enable quicker electronic medical record documentation, thus reducing the workload of clinicians while treating the patients (Kaufman et al., 2016). Yet, limited research has been done in the past in reducing psychiatrist workload using AI and NLP methods (Kazi & Kahanda, 2019). Choosing approriate data for training the machine learning model is important for predicting accurate results (Kazi & Kahanda, 2019). The past research has developed quality documentation tools for generating consistent and reliable data when rating physician case notes. For example, Physician Document Quality Instrument (PDQI-9) tool was used to assess the quality of physician case notes (admission, progress, and discharge summaries) (Stetson, Bakken, Wrenn, & Siegler, 2012). Similarly, the QNOTE instrument had high internal validity among subjects when subjects were asked to rate the clinical notes (Burke et al., 2014). However, no prior research has been conducted in developing a tool for delivering high inter-rater validity among subjects when annotating mental health transcripts. The focus of this exploratory study was to analyze the annotation process for mental health transcripts and develop a tool for transcript annotation. To authors' knowledge, this research is first of its kind to consider human factors to towards improving the annotation process for mental health transcripts.

METHODOLOGY

The study consisted of five synthetic mental health transcripts created by National Alliance on Mental Illness (NAMI) Montana. Each transcript was segmented into 90 (transcript A), 90 (transcript B), 100 (transcript C), 91 (transcript D), and 229 (transcript E) sentences respectively. Each sentence was annotated in six categories: Chief Complaint (CC), Medical History (MH), Family History (FH), Social History (SH), Client Detail (CD), and Other/Other Information (OT) (Kazi & Kahanda, 2019). These categories were suggested by the domain experts from NAMI. The subjects were asked to choose one of these six categories for each sentence which closely aligns with

the particular category type.

Clinician's Verification

Three clinicians were recruited in this study to evaluate the transcripts. Due to time constraints and limited availability of clinicians, the study only considered two transcripts for the verification process; transcript A and transcript B (Beck, Page, Buche, Rittman, & Gaiser, 2018; Satiani et al., 2018). The clinicians were asked to annotate transcript A and transcript B and asked to provide feedback for further improvements (Figure 1). The gold standard was achieved by comparing transcript A and transcript B of three participating clinicians. For example, if all three clinicians chose CC for a particular sentence, then the gold standard was CC for that sentence. If two clinician's choice matched and did not match with the third clinician, then the gold standard was chosen by the two clinicians with the same selection. For a particular sentence, if all the three clinician's selections are different, then the decision was determined by the clinician with the highest seniority. For example, for a particular sentence, if clinician 1, clinician 2, and clinician 3 chose CC, MH, and SH respectively, and if clinician 1 was the senior-most and had more experience than other two clinicians, CC was the gold standard because clinician 1 help more seniority.

Pilot Testing Phases

The pilot study was divided into two phases, phase 1 and phase 2 (Figure 1). The pilot study managed to recruit five subjects who completed both phase 1 and phase 2 and excluded other subjects who completed only one phase. Five transcripts were used in both phases; transcript A, transcript B, transcript C, transcript D, and transcript E.

In phase 1, before the transcript annotation, the subjects were given information about the focus of the study, definition of the categories, and four example sentences on how to correctly identify the category for each sentence. Phase 2 was conducted two weeks after phase 1 doing the same five transcripts. In addition, phase 2 also included a training transcript segmented into 90 sentences with six categories for each sentence. The subjects had to complete the training transcript before completing phase 2 transcripts. The training transcript guides the subject to choose the correct option if they selected the wrong category for that sentence.

Kappa Statistics for Measuring Agreement and Accuracy

Kappa statistics were used to measure the agreement among participating clinicians for transcript A and transcript B in clinician's verification (Fleiss, 1971; Nichols, Wisner, Cripe, & Gulabchand, 2010; Viera & Garrett, 2005). In pilot testing phases, to validate subject's performance after the training transcript, Kappa statistics were used to measure the subject's agreement and accuracy in both phase 1 and phase 2. Minitab (V19, Minitab LLC, State College, PA) was used in this study for the statistical analysis mentioned above.

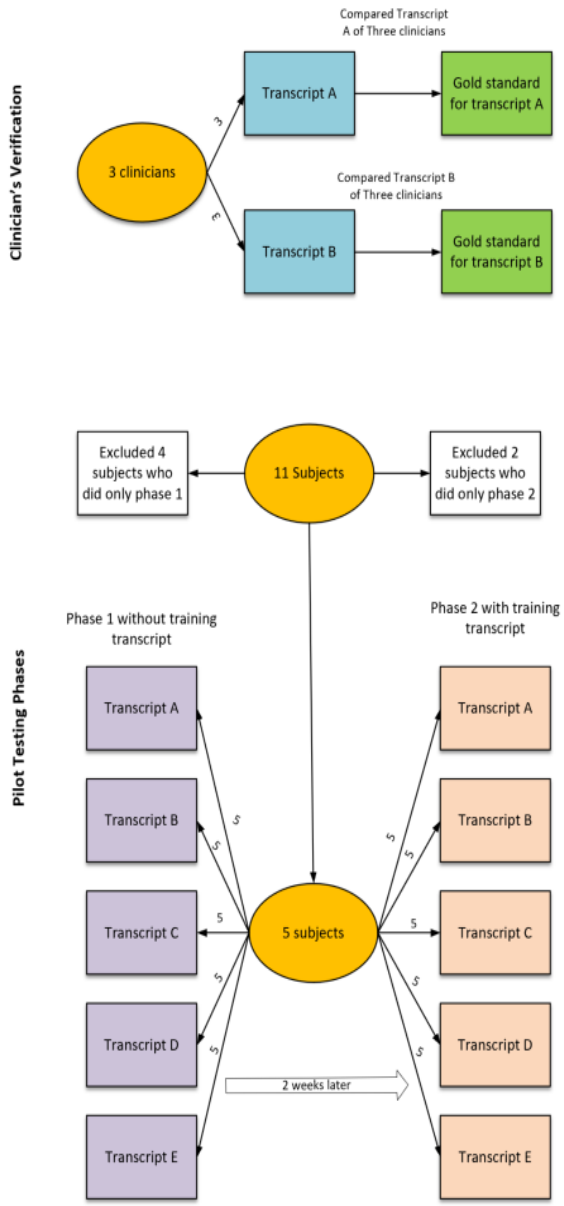

Figure 1- Experimental Design

RESULTS

Clinician's Verification Results

The clinician's verification results are shown in Table-1.

Table 1- Clinician results

	Transcript A (n=90)	Transcript B (n=90)
Percentage of sentences with complete agreement	35% (n=31)	71% (n=64)
Percentage of sentences with partial agreement	52% (n=47)	27% (n=24)
Percentage of sentences with no agreement	13% (n=12)	2% (n=2)
Fleiss Kappa Score (CI-95%)		
CC	0.26 (0.14 to 0.38)	0.59 (0.49 to 0.70)
CD	-0.05 (-0.17 to 0.07)	0.5 (0.38 to 0.62)
FH	0.41 (0.25 to 0.53)	0.19 (0.08 to 0.32)
MH	0.38 (0.26 to 0.5)	0.27 (0.15 to 0.39)
SH	0.26 (0.14 to 0.28)	0.69 (0.57 to 0.81)
OT	0.27 (0.15 to 0.39)	-0.02 (-0.15 to 0.57)
Overall	0.26 (0.19 to 0.33)	0.49 (0.42 to 0.57)

Consistency Results in Pilot Testing Phases

Comparing phase 1 and phase 2 (Table 2), except for transcript C and transcript E, the Kappa scores were higher in phase 2 than in phase 1. The transcript C and transcript E had lower Kappa scores may due to limited sample size of the subjects chosen in this study. The increase in Kappa score in phase 2 shows the higher inter-rater reliability among subjects after training transcript.

Accuracy Results in Pilot Testing Phases

To measure the accuracy of phase 1 and phase 2 results, transcript A of phase 1 and phase 2 were compared with transcript A of the gold standard (Figure 2). Transcript B of phase 1 and phase 2 were compared with transcript B of the gold standard (Figure 3). The Kappa scores denoted as k in Figure 2 and Figure 3 shows there was poor or no agreement between subjects and the gold standard. The student t-test was performed to compare the significant difference in accuracy percentage between phase 1 and phase 2 for transcript A and transcript B. The p-value for transcript A and transcript B were 0.04 and 0.10 respectively. In other words, the accuracy percentage after training was significantly higher in transcript A with an increase in mean accuracy from 24% to 47%. Similarly, the accuracy percentage after training was higher in transcript B with increase in mean accuracy from 63% to 79%.

DISCUSSION

Based on participating clinician's experience when annotating the transcripts, the clinicians perceived that transcript A to be more complex than transcript B. Transcript A was more complex due to patient type. In transcript A, the patient was more complex in terms of erratic behavior and responses leading to less cooperative and ultimately failure to comply with implicit and explicit rules for proper patienthood (Koekkoek, Hutschemaekers, van Meijel, & Schene, 2011; Koekkoek, van Meijel, & Hutschemaekers, 2006). The patient was experiencing multiple issues simultaneously like loneliness, separation from the significant other, suicide attempts, and single parenting pressure which lead to disruptive behavior and negative emotions towards doctors (Koekkoek, Van Meijel, Schene, & Hutschemaekers, 2009). This disruptive behavior in patients make the physicians hard to understand a patient's true problems and diagnose accordingly. The patient in transcript B was more corporative, straightforward, and going through one problem; germaphobia leading to anxiety. The study finds that the type of transcripts can affect inter-rater reliability among participating clinicians.

The participating clinicians also provided feedback on having additional categories specifically in medical history and mental history can directly help understand the patient problems for a proper diagnosis (Alegría et al., 2008; Mowbray, Oyserman, Bybee, & MacFarlane, 2002). Aforementioned, the current transcript annotation has six categories for each sentence. After consulting psychiatric experts from NAMI Montana and Chair of Psychiatric at the Billings Clinic-Montana, the authors updated the annotations by adding more categories related to mental health and medical history. The future transcript annotation will have the following categories Chief Complain (CC), Client Detail (CD), History of Present Illness (HPI), Past Psychiatric History (PPH), History of Substance Abuse (HSU), Family History (FH), Review of Systems (RS), and Others (OT). Note that the MH category was replaced by HPI, PPH, HSU, and RS categories. These new categories will likely deliver better categorization of transcript annotation resulting in an accurate patient diagnostic approach (Andreasen, Flaum, & Arndt, 1992; Mowbray et al., 2002).

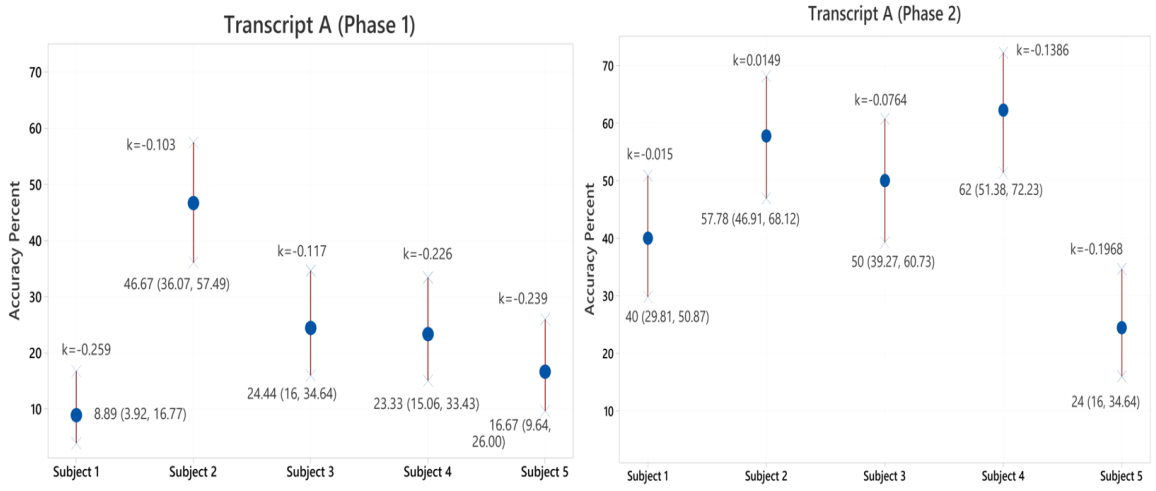

Figure 2- Kappa scores and accuracy percentage of subjects for Transcript A

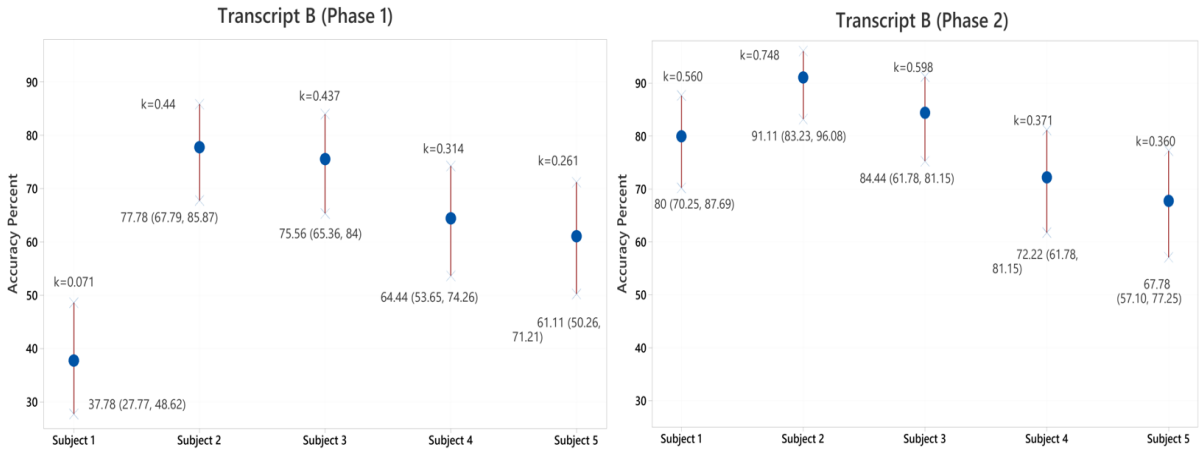

Figure 3- Kappa scores and accuracy percentage of subjects for Transcript B

In the clinician's verification phase, the participating clinicians' disagreement with each other was due to ambiguity of the sentence, and confusion between the categories. For example, in transcript A, sentence 7; "Is Carl is your boyfriend? Yeah" was annotated as SH by clinician 1, CC by clinician 2, and CD by clinician 3. Clinician 1 chose SH because sentence 7 was closely related to social history as per the definition. Clinician 2 comprehended that the patient was going through mental illness due to separation of the significant other (Carl); therefore, clinician 2 chose CC for that sentence. For sentence 7, clinician 3 chose CD, thinking CD was more appropriate than SH. In sentence 8; "How long have you been together? Couple of months" was annotated as SH by clinician 1 and clinician 2, and CD by clinician 3. Similarly, sentence 9; "We've been living together for the last four and a half weeks." was annotated as SH by clinician 1 and clinician 2, and CD by clinician 3. In both sentence 8 and sentence 9, we can notice the clinician 3 chose CD mistaking or confusing for SH. These examples emphasize the human factors on how each clinician perceives the transcript sentence and choose the category accordingly. This same pattern of clinicians' disagreement was also observed in transcript B. To improve the annotating experience and inter-rater reliability among participating clinicians in the future, the authors recommend the training transcript before starting the survey. The training transcript will explicitly illustrate the categories, eliminating ambiguousness and confusion when annotating. The training transcript will also help the participating clinicians in gaining experience in annotating complex transcripts (like transcript A) and deliver more standardized results.

This study explored the effects of training transcripts with subjects who are not clinicians. The subjects for pilot testing phases were selected randomly who had little or no prior knowledge of psychiatric nursing and mental health. Comparing the results of phase 1 and phase 2, it can be

inferred that introducing a training transcript before the survey can help improve the accuracy and consistency in the results. But the results from training transcript were not significantly reliable to use in the research. The reason for unreliable results after training transcripts was because the subjects did not have sufficient background knowledge in mental health and psychiatric nursing (Werner & Stawski, 2012). To obtain better results in data collection, authors recommend choosing subjects who have sufficient knowledge in mental health (such as psychiatric nursing students or clinician experts). The paper has limitations. First, there were not enough subjects in both expert's validation and pilot testing phases. Having more clinicians and more subjects in clinician's verification and pilot testing phases would have provided more clarity in the results. Second, the limited availability of clinicians resulted in considering only two transcripts for gold standard analysis. If the study had more gold standard transcripts, there might have been a better understanding of the subject's accuracy in pilot testing phases. Third, the study used synthetic transcripts in the expert's validation and pilot testing phases. There may have been different results and observations in the research if real patient transcripts replaced synthetic transcripts. Fourth, this is an exploratory study and our results may not be useful for all mental health transcripts.

CONCLUSION

This innovative study focuses on understanding the annotating process for mental health transcripts which will be applied in training machine learning models. Through this exploratory study, the research contributes to finding appropriate categories that should be included for transcripts annotations, suitable subject selection for data collections, and finally training the subjects before the survey for delivering high reliability in the data. Future research built from this study should include implementing the recommended data collection process to develop a novel system for psychiatrists, that combines automated case notes with an in-build treatment algorithm module.

COMPETING INTEREST

The authors have no competing interests to declare.

REFERENCES

Alegría, M., Nakash, O., Lapatin, S., Oddo, V., Gao, S., Lin, J., & Normand, S.-L. (2008). How missing information in diagnosis can lead to disparities in the clinical encounter. *Journal of public health management and practice: JPHMP, 14*(Suppl), S26.

Andreasen, N. C., Flaum, M., & Arndt, S. (1992). The Comprehensive Assessment of Symptoms and History (CASH): an instrument for assessing diagnosis and psychopathology. *Archives of general psychiatry, 49*(8), 615-623.

Beck, A. J., Page, C., Buche, J., Rittman, D., & Gaiser, M. (2018). Mapping Supply of the US Psychiatric Workforce.

Burke, H. B., Hoang, A., Becher, D., Fontelo, P., Liu, F., Stephens, M., . . . Baxi, N. S. (2014). QNOTE: an instrument for measuring the quality of EHR clinical notes. *Journal of the American Medical Informatics Association, 21*(5), 910-916.

Butryn, T., Bryant, L., Marchionni, C., & Sholevar, F. (2017). The shortage of psychiatrists and other mental health providers: causes, current state, and potential solutions. *International Journal of Academic Medicine, 3*(1), 5.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin, 76*(5), 378.

Kaufman, D. R., Sheehan, B., Stetson, P., Bhatt, A. R., Field, A. I., Patel, C., & Maisel, J. M. (2016). Natural language processing–enabled and conventional data capture methods for input to electronic health records: a comparative usability study. *JMIR medical informatics, 4*(4), e35.

Kazi, N., & Kahanda, I. (2019). *Automatically Generating Psychiatric Case Notes From Digital Transcripts of Doctor-Patient Conversations.* Paper presented at the Proceedings of the 2nd Clinical Natural Language Processing Workshop.

Koekkoek, B., Hutschemaekers, G., van Meijel, B., & Schene, A. (2011). How do patients come to be seen as 'difficult'?: A mixed-methods study in community mental health care. *Social science & medicine, 72*(4), 504-512.

Koekkoek, B., van Meijel, B., & Hutschemaekers, G. (2006). " Difficult patients" in mental health care: a review. *Psychiatric Services, 57*(6), 795-802.

Koekkoek, B., Van Meijel, B., Schene, A., & Hutschemaekers, G. (2009). Problems in psychiatric care of 'difficult patients': a Delphi-study. *Epidemiology and Psychiatric Sciences, 18*(4), 323-330.

McCluskey, P. D. (2019). Physician Burnout Now Essentially a Public Health Crisis. *Boston Globe*.

Mowbray, C., Oyserman, D., Bybee, D., & MacFarlane, P. (2002). Parenting of mothers with a serious mental illness: Differential effects of diagnosis, clinical history, and other mental health variables. *Social Work Research, 26*(4), 225-240.

Nichols, T. R., Wisner, P. M., Cripe, G., & Gulabchand, L. (2010). Putting the kappa statistic to use. *The Quality Assurance Journal, 13*(3-4), 57-61.

Satiani, A., Niedermier, J., Satiani, B., & Svendsen, D. P. (2018). Projected workforce of psychiatrists in the United States: a population analysis. *Psychiatric Services, 69*(6), 710-713.

Stetson, P. D., Bakken, S., Wrenn, J. O., & Siegler, E. L. (2012). Assessing electronic note quality using the physician documentation quality instrument (PDQI-9). *Applied clinical informatics, 3*(02), 164-174.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Fam med, 37*(5), 360-363.

Werner, S., & Stawski, M. (2012). Mental health: Knowledge, attitudes and training of professionals on dual diagnosis of intellectual disability and psychiatric disorder. *Journal of Intellectual Disability Research, 56*(3), 291-304.

APPENDIX D

WIP: DETECTION OF STUDENT MISCONCEPTIONS OF ELECTRICAL CIRCUIT CONCEPTS IN A SHORT ANSWER QUESTION USING NLP

2021 ASEE ANNUAL CONFERENCE
Virtual Meeting | July 26–29, 2021 | Pacific Daylight Time
Paper ID #32929

WIP: Detection of Student Misconceptions of Electrical Circuit Concepts in a Short Answer Question Using NLP

Prof. James P Becker, Montana State University, Bozeman

James Becker is a Professor of electrical and computer engineering at Montana State University. His professional interests include microwave circuits, radio frequency electronics, nanoelectronics, pedagogical research, and distance education.

Dr. Indika Kahanda, University of North Florida

Dr. Indika Kahanda is an Assistant Professor in the School of Computing at the University of North Florida, where he directs the bioinformatics, biomedical informatics and medical informatics lab. Prior to that Dr. Kahanda worked as an Assistant Professor in the Gianforte School of Computing at Montana State University. He received his Ph.D. in Computer Science from Colorado State University in 2016 in the area of Bioinformatics, a Master of Science in Computer Engineering from Purdue University in 2010, and a Bachelor of Science in Computer Engineering from University of Peradeniya, Sri Lanka in 2007.

Nazmul H. Kazi, Montana State University

Nazmul Kazi is a master's student of Computer Science at Montana State University. His research interests include the application of Artificial Intelligence, Deep Learning, Natural Language Processing, and Parallel Computing.

WIP:

Detection of Student Misconceptions of Electrical Circuit Concepts in a Short Answer Question Using Natural Language Processing

Abstract

While the use of writing exercises in gateway STEM courses that focus on solving numeric problems is not widespread, there is evidence that students could benefit from the addition of such exercises [1]. Writing exercises may be effective in both uncovering student misconceptions that are not necessarily apparent with typical computation problems, and as tools to foster conceptual change and metacognitive skill.

In this paper, pilot studies of the use of two Natural Language Processing (NLP) techniques to identify common misconceptions in the writing of students in a course on electric circuit analysis are described. Performance on the writing exercise in question has been shown to correlate with a student's performance in the course [2]. This is of particular interest as the writing exercise has been administered during the fifth class period, sufficiently early to direct additional resources to the success of students appearing to be at-risk for failing the course. Realizing an automated software solution to analyze the responses to this exercise would remove burden on instructor time and open the door to immediate and personalized feedback to the student.

The first pilot study was run to determine how successful a simplistic rule-based approach would be in identifying the most common misconceptions found in a writing exercise requiring a student to speculate on the change in the power in the elements of a resistive circuit with a change to a single resistor value. An open-source NLP rule-based matching engine within spaCy [3] was used. The corpus consisted of one hundred and eighty-five unique responses to the question. Precision, recall, and F1-score [4] were used to assess the effectiveness of the rule-based NLP pipeline in comparison to that of a subject matter expert in identifying responses exemplifying seven misconceptions. Should this NLP pipeline be used in a system in which feedback is to be given to the student, a Directed Line of Reasoning (DLR) approach [5] would be beneficial in cases in which identification of a given misconception is in doubt. Considering this pilot study employed an extremely simplistic purely lexical-level rule-based classifier, the results are very promising and suggest the planned approach of developing a highly accurate, advanced rule-based classifier encompassing lexical/syntax/semantic driven rules is viable. As a compliment to the rule-based approach, this paper also describes a pilot study of the use of BERT (Bidirectional Encoder Representations from Transformers) [6], a machine learning approach that has shown tremendous promise in short-answer grading [7].

I. Introduction

Student struggles in gateway STEM courses may arise from a variety of factors. Two commonly identified impediments to student success in such courses include inadequate "prerequisite learning and thinking skills" [8] and the inaccurate prior knowledge [9]. Learning and thinking skills fall under the terms "self-regulated learning" [10] and "metacognition" [11] and involve

skills such as accurate self-assessment of knowledge ("knowledge of cognition") and the ability to formulate and follow a plan for mastery ("regulation of cognition"). For mastery of the content studied in gateway courses, students must often demonstrate an understanding of concepts for which they may come to class with inaccurate models and frameworks. For example, it has been found that students entering courses on statistics often have intuitions regarding probability and statistics that are at odds with accepted reasoning [12]. Such erroneous mental frameworks and misconceptions are often challenging to correct [13].

There are many identified misconceptions among students beginning courses on electric circuit analysis [14]-[20]. For example, it is not uncommon for a student to fail to appreciate that an electric circuit is a system and rather believe that the "flow" of current is sequential. Colloquial phrasings such as "current flow" are often used by instructors for expediency but may strengthen student misconceptions. While evidence of certain misconceptions held by students in the electric circuit domain may be found in their computations, student writings offer the most vivid insight into a student's thinking. The value of writing as a tool for uncovering a student's misconceptions has been noted in other disciplines such as the medical field [21]. Unfortunately, grading and providing feedback to students on their written work is time consuming. This burden on instructor time may be a factor why, beyond common written works such as laboratory reports, courses such as electric circuit analysis or statics and dynamics are almost exclusively computation based. The authors of this paper do not suggest eliminating computation problems in gateway STEM courses, but rather to complement such problems with conceptual writing exercises as such exercises may be the key to effecting conceptual change particularly in the case of robust misconceptions.

The remainder of this paper focuses on a describing the results of pilot studies in the use of two techniques in natural language processing (NLP) to identify misconceptions in the responses of students to a writing quiz in an introductory circuits course, EELE 201, at Montana State University. The details of the writing quiz may be found elsewhere [2],[22]. In short, the question refers to a four-element circuit as depicted in Figure 1 and asks the student to argue what will happen to the power (increase, decrease, or remain the same) of each of the circuit's four elements, when the resistance of resistor R_2 decreases. Students are told to treat all elements as ideal including the independent voltage source and to thoroughly justify their responses.

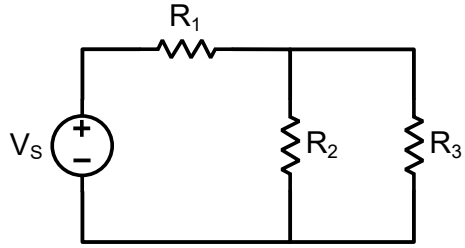

Figure 1: Circuit schematic diagram of the circuit students are to consider in a writing quiz.

This written quiz has been offered on the fifth period of class in EELE 201 and student performance on the quiz has been found to correlate with their performance in the class, making the writing exercise of significant value in identifying students likely to struggle in the course. It should be noted that prior to EELE 201, students do receive minor exposure to KVL, KCL, Ohm's Law, and series and parallel resistor combinations in a freshman course meant to introduce students to the electrical and computer engineering majors. The ultimate goal for this and other writing exercises used in EELE 201 is to embed a given writing exercise within a web-based application

that without human intervention, evaluates a student's response and provides meaningful feedback to strengthen the student's conceptual understanding while promoting active metacognition.

II. The Rule-Based Matching Approach

While machine learning methods are common in natural language processing (NLP) applications [6],[7], they are most powerful when a large, labeled corpus is available and the task is simply to classify the responses, for example, to grade them. As the existing corpus of responses is modest (185) and a key goal in the current work is to identify examples of misconceptions in students' writing, we elected to begin by using a hand-crafted rule-based approach. The matching engine, "matcher", within the open-source NLP framework spaCy [3] was used. We divided the 185 responses into two groups, a training set consisting of 60% of the responses and a test set consisting of the remaining 40%. The training set was considered when generating rules to identify a given misconception. Each rule consisted of a search for an ordered set of between two and four key words, together which likely indicated a given misconception. Once the rules were established, they were applied to each response within the test set on a sentence-by-sentence basis. Measures of precision, recall, and F1-score were determined at the response level.

The *precision* of our rule-based NLP classifiers is the fraction of correct predictions of responses exemplifying a given misconception over the number of instances in which the classifier identified a response as exemplifying the considered misconception. The *recall* of a given classifier is the fraction of accurate identifications of a response exemplifying the misconception considered over the total number of responses that were tagged by the expert to exemplify the misconception. From measures of the precision and recall, the F1-score [4] was computed.

A. Sequential Misconception

A sequential misconception in terms of electric circuits is one in which it is believed that elements that are further "downstream" from a source (such as R_2 and R_3 in the example circuit of Figure 1) "receive" current after elements closer to the source (R_1 in the example circuit). With such a misconception, it is likely that a student will think that changes in R_2 have no effect on the potential difference and current associated with R_1 or V_s. The following two examples are drawn from the corpus of responses and indicative of a sequential misconception on the part of the writer.

Since R_1 is a component lying before R_2, its power absorbed should be unchanged.

The power at R_1 should be unaffected by changing R_2. R_1 is not affected as its current does not depend on R_2.

In this work, examples such as the first that refer to the relative locations of R_1 and R_2 to argue that the power in R_1 will not change with a decrease in the value of R_2 are termed "explicit" examples of the sequential misconception. Such examples are easily identified using rules looking for two-word combinations such as "before R_1" and "after R_2". Examples such as the second appear to imply the sequential misconception and so are termed "implied" examples. Extracting such examples typically required a three-word combination with the words separated by other "tokens". For example, the second example provided could be extracted using a rule looking for

the word sequence "R_1 not affected" but allow another word or words to exist between "R1", "not", and "affected". Tokenization is one of the first steps in an NLP pipeline and often includes more complex tasks such as stemming or lemmatizaton [23]. In this work a basic tokenization scheme was followed, namely, separating words by spaces and punctuation marks; in some instances the lemma (i.e. base form) of a given word was invoked to make a given rule more widely applicable. While extracting explicit examples of the sequential misconception was straightforward, attempting to identify the implied cases of the sequential misconception resulted in some false positives. In this case, a false positive is a response identified by the rule-based algorithm as a sequential misconception that was not tagged as a sequential misconception by the content expert. A set of rules were developed considering the training set; the rules were then applied to the test set. As identified by the content expert, approximately 16% of the 185 responses considered included evidence of either explicit or implicit sequential thinking.

The precision, recall, and F1-score were found to be 0.632, 1, and 0.77, respectively. The value of unity for the recall indicates that the rules were sufficient to extract <u>all</u> instances in the unseen test set tagged by the content expert as exemplifying either an explicit or implicit sequential misconception. Of the seven false positives, four exemplified another misconception, what in this work is termed a "localized" misconception. Students exemplifying this misconception express the belief that only quantities associated with R_2, such as its current or power would change. All three of the remaining false positives indicated that they believed the power associated with R_1 either increased or decreased, conclusions that are in opposition to the sequential misconception. As described in [22], a web-based application for the writing question has been created though it has yet to be powered with NLP. One of the features of the existing application is a drop-down selection for each of the four circuit elements for which the user must indicate whether their response supports that the power associated with a given element *increases*, *decreases*, or *stays the same*. The original intent of including the drop-down was simply to remind students they were to address the power associated with each element as it was observed from the first deployment of the handwritten quiz that many students failed to address all four elements. Since we will have unambiguous information regarding what a student feels are the correct answers, specific rules based on a student's drop-down selections can be applied once certain responses are removed from consideration. In this case, the rules would be applied after the three responses indicating a change in the power of R_1 were removed. So too, responses that correspond to drop-down selections in which ONLY the power of R_2 was believed to change could be removed to avoid lumping localized with sequential misconceptions. Even without using drop-down selection to eliminate the potential for false positives, the results are promising considering only a very simplistic word-matching approach was used.

B. Constant Voltage Errors

The second most common error that showed up in in student responses, found in approximately 15% of responses, was the belief that the voltage drop across R_2 and/or R_3 did not change as the resistance of R_2 decreased. Students making this error often did not justify why they felt the voltage drop across one or both of these resistors would not change as the value of R_2 decreased. Perhaps these students recalled the fact that the potential difference across the ideal voltage source would not change and erroneously extended the thought to the other components, failing to appreciate how the potential difference redistributes across the resistors with changes in the value

of R_2. An example response follows. Note that in addition to the constant voltage error, the given response suggests localized thinking on the part of the student.

The total resistance is found by combining R_2 and R_3 in parallel and adding R_1. V_s does nothing different. R_1 has the same potential difference as before. R_2 is less resistive but the potential difference has to be the same so to make up for the loss in resistance. The current is stronger therefore the power related to R_2 increases due to the stronger current and the same voltage. Power equals voltage times current. R_3 remains the same.

Once again, a series of rules were generated based on an analysis of the training set and implemented using matcher. The rules were then applied to the unseen test set. The precision, recall, and F1-score were found to be 0.69, 0.82, and 0.75, respectively. In examining the false positives, it was found that augmenting the basic capability of matcher with a function that excluded a match if there existed the term "voltage source" in the tagged sentence, the precision could be increased to one. The false negatives were found to be cases in which evidence for the misconception was implied as suggested in the following example.

… but R_3 is not directly affected in any way by R_2, thus I_3 stays the same. If I_3 remains the same then the power must also remain the same.

In the above example, specific mention of the potential difference across R_3 is not made. The example was tagged by the content expert with the constant voltage error based on the implication of having I_3 and the power of associated with R_3 being unchanged.

C. Misconception with an Ideal Independent Voltage Source

In a typical course on electric circuit analysis, students are introduced to the model of an ideal independent voltage source. Ideal independent voltage sources provide a constant potential difference across their terminals. The current and thus the power associated with such a source depend on the circuit to which the source is connected. For simplification purposes, batteries are often modeled as ideal sources, though the limitations of such a model for batteries should be explained. As has been noted elsewhere [14],[15] a common misconception regarding batteries (independent voltage sources) is that they are sources of constant current. This misconception, along with the notion that the power associated with a voltage source is constant turned up in approximately 12% of the responses. Two examples follow.

Vs would stay the same power wise because it is independent from the other elements of the circuit.

Given that the voltage source remains unchanged, the voltage and current through it remain the same, so the power does not change.

Considering the training set, rules were developed to catch misconceptions regarding the ideal independent voltage source. Again, these rules were based on either a two- or three-word set in order, with tokens allowed between the key words. The precision, recall, and F1-score were found to be 0.50, 1.0, and 0.67, respectively. Therefore, while the rules caught all responses tagged by the content expert to suggest an error in applying the concept of an ideal independent source, the rules picked out an equal number of false positives. Seventy percent of the false positives could

be ascribed to another misconception. If knowledge from the use of drop-down selections as previously described were employed, the remaining 30% of the false positives would not occur, thus improving the performance (precision = 0.59, recall = 1.0, F1 = 0.74) of the simple matching algorithm. It is interesting to note that considering the full corpus (185 responses), the terms "ideal source" or "independent source" when referring to V_s showed up in just over 8% of the responses. Of the responses specifically identifying V_s as an ideal or independent source, nearly half (~47%) misused the term.

D. Resistor Combination Errors

In responding to the writing quiz, students scoring at the upper end of the scale recognized that thinking about the equivalent resistance "seen" by the source was an effective starting point. Such students identified that R_2 and R_3 are in parallel and this parallel combination is in series with R_1. From there, successful students would note that in decreasing R_2, the effective resistance seen by the source decreases. Approximately 9% of the responses in the corpus of 185 unique answers revealed at least one error regarding claims made in terms of resistance; three examples follow.

R_1 and R_3 are in parallel therefore...

...if R_2 goes to zero, then R_1 and R_3 are in series

The equivalent resistance of the circuit increases as R_2 decreases.

The first response demonstrates a clear misunderstanding of what is required for resistors to be in parallel. The error in the second statement likely stems from a misunderstanding of the result of placing a short circuit in parallel with a resistance. The third statement reveals a lack of awareness that the net resistance of two resistors in parallel is less than the smaller of the constituent resistors. It should be remembered that the writing quiz was given early in the semester (fifth class period); nevertheless, promptly correcting such errors is critical. Once again rules to pick out resistance combination errors were composed based on an examination of the training set (60% of the corpus). The rules were then applied to the test set (remaining 40% of the corpus) to establish values of precision (1.0), recall (0.71) and F1-score (0.83). It is interesting to note that the precision was perfect, indicating that zero false positives were identified. It appears then that one would not have to invoke a specific set of rules using knowledge of the drop-down selections when attempting to identify resistance combination errors. That fact that recall was not perfect reflects that additional rules are necessary to capture all the examples of resistor combination errors in the existing corpus. The following is a somewhat obscure example of a "resistance combination error" that was not caught by the rules created when considering the training set.

The equivalent resistance must stay constant, so as the resistance of R_2 decreases, the resistance of R_3 must increase.

An advantage of rule-based systems is that the rules can grow without substantial change in the system. Considering this most recent example, a simple rule can be added to catch such a rare misconception (1 of 185 responses suggested that the equivalent resistance must remain constant).

E. Localized Misconception

As noted in the discussion of the sequential misconception, in terms of the work here, a localized misconception is one in which the responder believes that for a quantity such as current or power to change in an element, that element must itself change. In the problem considered, only the resistance value of R_2 changes and so a student falling prey to the localized misconception would express that only the power of R_2 would change. Just over 4% of the responses were found to exhibit evidence of the localized misconception. In considering the training set when developing the rules to catch the localized misconception, it was noticed that our chosen procedure of analyzing responses one sentence at a time, while effective for the other misconceptions noted, would not be in the case of the localized misconception. The following example illustrates the need for a somewhat more sophisticated approach.

The power associated with Vs with the resistance of R_2 decreasing will be the same. The power of R_1 will remain the same as the current and resistance stay the same. The power of R_2 will decrease for as R decreases, so will power. The power of R_3 will remain the same as current and resistance remain the same.

To deduce that the student exhibits the "localized" misconception, three of the four sentences (sentence 1, 2, and 4) would need to be tagged, as individually they show the author thinks that the power of V_s, R_1, and R_3 will remain the same, respectively. While certainly this could be done, recalling that the existing web-based application has drop-down selections to indicate the nature of a change in power of each of the elements, any response that indicated only the power of R_2 changed could be identified as a localized misconception. Such an approach is the most efficient means to tag responses for the localized misconception in the given question.

F. Precedence of Current Misconception

Consider the following response from the corpus.

If the equivalent resistance for this circuit decreases than [sic] the current will increase in order for the voltage to remain constant.

A common misconception regarding electric circuits has to do with the relationship between potential difference and current. While the correct understanding is to appreciate that it is the potential difference that causes current (i.e. the flow of charge), many operate under the erroneous idea that current causes potential difference. Such a misconception may lead to errors in applying Ohm's law for example, as well as not appreciating that a potential difference can exist across an open circuit [24]. Just under 4% of responses provided direct evidence of this misconception. Rules to catch this misconception were created based on the training set and found to successfully capture all instances of the misconception in the test set with zero false positives. This suggests that perfect precision, recall, and F1-score were achieved. The number of instances of this misconception (4 in the training set and 3 in the test set) was so small, that the results should only be taken to suggest that capturing examples of this misconception may be done relatively easily with the chosen approach.

G. Conservation of Energy Errors

A search was performed to identify all responses that included either the term *conservation of energy* or *conservation of power* and their variants. Of the 185 responses, fourteen (~8%) included such a reference. In ten of these fourteen, the use of the term is either incorrect or ambiguous. Consider the following examples.

The power associated with V_s and R_1 will stay the same because Vs is the only provider of power in the circuit and because power is conserved it must stay the same.

I know that R_2 and R_3 are connected in parallel so they won't have the same currents. But, I would need to combine all my resistors first to find the total current. After finding all the currents, figure out [sic] power equations before and after R_2 (keeping in mind that power is supposed to be conserved but in this case, it's not).

The first statement attempts to use the notion of conservation of power to argue (incorrectly) that the power of the voltage source must remain constant as it is the only source and thus only provider of energy in the circuit. In the second case, while mentioning that, "power is supposed to be conserved," the student believes he/she has found an example in which it is not. Students who used conservation of energy correctly considered the power (energy per unit time) of each element in the circuit before coming to a conclusion. As the method pursued in this pilot study looked only at the sentence level and did not try to tie the meaning from one sentence in a response to another within the response, no attempt was made to determine precision, recall, and F1-score. It is quite possible that values for those quantities would be considerably higher than warranted due to the fact that it was trivial to extract examples of the use of the concept of conservation of energy and that most of the uses were not applied properly.

III. A Machine Learning Approach

As a compliment to the rule-based approach, a machine learning approach was investigated. We modeled the task of predicting responses with sequential misconceptions as a supervised binary classification problem in which examples are the whole responses and the labels are positive if the response contains a sentence conveying the misconception. We used a BERT model pre-trained on a large corpus of general English text data in a self-supervised fashion. Then, we fine-tuned the BERT model for our task using the labeled training data. We used the exact same data and experimental setup used to evaluate the rule-based approach; a comparison of key metrics between the rule-based and BERT approaches are provided in Table 1.

Table 1: Comparison of models for sequential misconception prediction

Model	Precision	Recall	F1
Rule-based	0.63	1.0	0.77
BERT	0.90	0.75	0.82

As depicted in the table, BERT's precision is excellent but its recall, which we claim is more important for misconception prediction, is relatively lower than that of the rule-based approach. With more training data and domain-specific pre-training [25], data-hungry models such as BERT

will likely outperform rule-based models in both metrics. For our purposes, a reasonable approach in developing a misconception detection system for a given conceptual-based writing problem would be to create a combined system in which the rule-based method is initially active until sufficient gold-standard data is generated from multiple offerings of the quiz at which point the BERT model may start assisting for improved accuracy. Regardless of the chosen NLP approach to misconception detection, the ultimate aim is to correct the misconception through some form of feedback. In the following section, the proposed method of feedback for the conceptual-based writing problem is described and demonstrates that perfect interpretation of the student's response is not necessary.

IV. Correcting the Misconceptions through Feedback

The original intent of the writing quiz described above was to identify students at-risk to fail the course [2]. Currently, additional writing quizzes are being explored to correct student misconceptions and to foster metacognitive skill. The manner in which students are given feedback is no doubt crucial. Our conception of feedback in terms of the writing quizzes is not primarily one of simply articulating correctness or incorrectness of a response but rather is itself a form of instruction as suggested in [26]. As noted by Hattie [27], feedback, "is most powerful when it addresses faulty interpretations, not a total lack of understanding. Under the latter circumstance, it may even be threatening to a student." Hattie goes on to note that, "a key theme arising from this review of the literature is the importance of ensuring that feedback is targeted at students at the appropriate level." Simply describing to the class as a whole, a correct approach to answering the question is therefore not the most effective for it does not target the individual. Our desire is to use a web-based application to evaluate student responses and to immediately provide them individualized feedback based on a Directed Line of Reasoning (DLR) algorithm [5].

In the previous sections, we have described how employing drop-down selections in the web application and simple word matching using an open-source tool, or using a machine learning approach such as BERT, we can begin to assemble a model of a given student in terms of their understanding of the writing quiz. Based on any identified misconceptions, or solid conceptions such as the value of considering the equivalent resistance, a starting point in the DLR could initiate. For example, should a student mention conservation of energy, the algorithm could congratulate the student on considering the term in the context of the problem and then lead them through simple question and answering (perhaps through multiple choice selections) to a proper understanding of the term. In a similar fashion, should a response be tagged with a sequential misconception, the DLR algorithm could help the student come to appreciate the importance of looking at a circuit as a system. The key point with such an approach is that it can provide a level of individualized feedback that is immediate as opposed to either a class discussion (not individualized) or instructor feedback given days later (not immediate) due to grading demands.

V. Conclusions

As a means to identify misconceptions with regard to basic concepts related to electric circuit analysis within the writings of electrical and computer engineering students, a simple word matching approach applied at the sentence level, has been described. Even though a very rudimentary NLP approach was implemented, the results have shown considerable promise for the task. For example, the precision, recall, and F1-score were 0.632, 1.0, and 0.77, respectively when

searching for evidence of a sequential misconceptions. Other misconceptions and errors, such as that associated with resistor combinations were found with similar metrics. Misconceptions such as "localized thinking" would not be effectively found by analyzing the word choices of single sentences, but rather need a more sophisticated algorithm that looks at the meaning (as revealed through word search) of multiple sentences within a response. More simply, using the drop-down selections of the existing web-based application would permit the localized misconception to be identified without resorting to an NLP algorithm altogether. Preliminary results from BERT, a machine learning approach apparently well-suited to short answer evaluation, were described. The value of such an approach grows as the corpus of responses used for training grows. Whether the evaluation is done by a human or via an NLP-based algorithm as described, there is often ambiguity for the very reason that students were often found to fail to adequately justify their responses to the considered conceptual writing quiz. This is where a Directed Line of Reasoning approach to providing feedback would be most useful.

Acknowledgements

This material is based upon work supported of the National Science Foundation under Grant No. 1504880. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Works Cited

[1] Writing to Develop Mathematical Understanding, David K. Pugalee, Christopher-Gordon Publishers, Inc. Norwood, Massachusetts, p. 111, 2005.

[2] Becker, J. P., & Plumb, C. (2018, June), Board 8: "Identifying At-Risk Students in a Basic Electric Circuits Course Using Instruments to Probe Students' Conceptual Understanding," paper presented at 2018 ASEE Annual Conference & Exposition, Salt Lake City, Utah. 10.18260/1-2—30110.

[3] https://spacy.io/usage/rule-based-matching; last accessed 12/17/2020

[4] D.M.W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation", Journal of Machine Learning Technologies, Vol. 2, Issue 1, pp. 37-63, 2001.

[5] Bruce Mills, Martha Evens, Reva Freedman, "Implementing Directed Lines of Reasoning in an Intelligent Tutoring System Using the Atlas Planning Environment", IEEE Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04).

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (2019) Association for Computational Linguistics; Minneapolis, Minnesota.

[7] Sung C., Dhamecha, T.I., Mukhi, N. (2019) Improving Short Answer Grading Using Transformer-Based Pre-training. In: Isotani S., Millán E., Ogan A., Hastings P., McLaren B., Luckin R. (eds) Artificial Intelligence in Education. AIED 2019. Lecture Notes in Computer Science, vol 11625. Springer, Cham.

[8] Robert A. Blanc, Larry E DeBuhr and Deanna C. Martin, "Breaking the Attrition Cycle, The Effects of Supplemental Instruction on Undergraduate Performance and Attrition," Journal of Higher Education, Volume 54, Number 1, 1983, pp. 80-90.

[9] Susan Ambrose, et al. "How Does Students' Prior Knowledge Affect Their Learning?" in How Learning Works: Seven Research-Based Principles for Smart Teaching, John Wiley & Sons, pp. 10-39, 2010.

[10] Deborah L. Butler and Philip H. Winne, "Feedback and Self-Regulated Learning: A Theoretical Synthesis," Review of Educational Research, Vol. 65, No. 3, pp. 245-281, 1995.

[11] John Flavell, "Metacognitive Aspects of Problem Solving," in The Nature of Intelligence, Lauren B. Resnick ed., Lawrence Erlbaum Associates, Hilsdale, N.J., 1976.

[12] Clifford Konold (1995) Issues in Assessing Conceptual Understanding in Probability and Statistics, Journal of Statistics Education, 3:1, , DOI: 10.1080/10691898.1995.11910479.

[13] Michelene T.H. Chi, "Commonsense Conceptions of Emergent Processes: Why Some Misconceptions Are Robust," Journal of the Learning Sciences, 14(2), pp. 161-199, 2005.

[14] Paula Vetter Engelhardt and Robert J. Beichner, "Students' Understanding of Direct Current Resistive Electrical Circuits," American Journal of Physics, Vol. 72 (98), pp. 98-115, 2004.

[15] Tatiana V. Goris and Michael J. Dyrenfurth, "How Electrical Engineering Technology Students Understand Concepts of Electricity. Comparison of Misconceptions of Freshmen, Sophomores, and Seniors," Proceedings of the 2013 American Society for Engineering Education Annual Conference and Exposition. Paper ID 5849.

[16] David P. Tallant, "A Review of Misconceptions of Electricity and Electrical Circuits," The Proceedings of the Third International Seminar on Misconceptions and Educational Strategies in Science and Mathematics, August 1-4, 1993.

[17] Deepika Sangam and Brent K. Jesiek, "Conceptual Understanding of Resistive Electric Circuits Among First-Year Engineering Students," Proceedings of the 2012 American Society for Engineering Education Annual Conference and Exposition.

[18] Richard Gunstone, Brian McKittrick, Pamela Mulhall, "Textbooks and their authors: another perspective on the difficulties of teaching and learning electricity," in *Research and the Quality of Science Education*, Kerst Boersma, Martin Goedhart, Onno de Jong, Harrie Eijkelhof,.eds. Springer, 2005.

[19] Deepika Sangam and Brent K. Jesiek, "Conceptual Gaps in Circuits Textbooks: A Comparative Study," IEEE Transactions on Education, Vol. 58, August 2015, pp. 194-202.

[20] Brian J Skromme, "Addressing Barriers to Learning in Linear Circuit Analysis," 122nd ASEE Annual Conference and Exposition, Paper ID #14125, June 2015.

[21] Olde Bekkink et al., "Uncovering students' misconceptions by assessment of their written questions," BMC Medical Education (2016) 16:221 DOI 10.1186/s12909-016-0739-5

[22] Becker, J. P., & Sior, E., & Hoy, J., & Kahanda, I. (2019, June), Board 11: "Predicting At-Risk Students in a Circuit Analysis Course Using Supervised Machine Learning," presented at 2019 ASEE Annual Conference & Exposition, Tampa, Florida. 10.18260/1-2--32185

[23] *Natural Language Processing in Action*, Hobson Lab, Cole Howard, and Hannes Hapke, Manning Publications, 2018.

[24] R. Cohen, B. Eylon, and U. Ganiel, "Potential difference and current in simple electric circuits: A study of students' concepts," American Journal of Physics 51, 407 (1983)

[25] Chul Sung, Tejas Dhamecha, Swarnadeep Saha, Tenghai Ma, Vinay Reddy, and Rishi Arora, "Pre-Training BERT on Domain Resources for Short Answer Grading," Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 6071-6075.

[26] R.W. Kulhavy, "Feedback in written instruction." Review of Educational Research, 47(1), 211–232 (1977).

[27] John Hattie and Helen Timperley, "The Power of Feedback," Review of Educational Research March 2007, Vol. 77, No. 1, pp. 81–112.