Mining the Clinical Narrative: All Text Are Not Equal

Keith Feldman
Dept. of Computer Science
and Engineering, and iCeNSA
University of Notre Dame
Notre Dame, IN 46656
Email: kfeldman@nd.edu

Nicholas Hazekamp
Dept. of Computer Science
and Engineering
University of Notre Dame
Notre Dame, IN 46656
Email: nhazekam@nd.edu

Nitesh V. Chawla
Dept. of Computer Science
and Engineering, and iCeNSA
University of Notre Dame
Notre Dame, IN 46656
Email: nchawla@nd.edu

***Abstract*—Over the past decade, the application of data science techniques to clinical data has allowed practitioners and researchers to develop a sundry of analytical models. These models have traditionally relied on structured data drawn from Electronic Medical Records (EMR). Yet, a large portion of EMR data remains unstructured, primarily held within clinical notes. While recent work has produced techniques for extracting structured features from unstructured text, this work generally operates under the untested assumption that all clinical text can be processed in a similar manner. This paper provides what we believe to be the first comprehensive evaluation of the differences between four major sources of clinical text, providing an evaluation of the structural, linguistic, and topical differences among notes of each category. Our conclusions support the premise that tools designed to extract structured data from clinical text must account for the categories of text they process.**

I. Introduction

The past decade has held witness to a significant transformation of the United States healthcare industry. These changes have affected every part of the healthcare spectrum, from the financial organization to the daily activity of medical professionals. One of the most prominent changes has been the emergence and implementation of electronic medical records (EMRs). Amidst government mandates and an increased demand for collaboration between medical institutions, EMRs have become a standard in clinical practice. The utilization of *Big Data* aggregated from millions of clinical EMRs has sparked the growth of a new field of research known as healthcare informatics, blending together the statistical foundations of data mining and machine learning, with the clinical outcomes of traditional medical research.

Today, the emergence of healthcare informatics has enabled medical research to push further, transitioning from preventative care to personalized medicine. Through the use of analytics researchers have been able to provide medical insights for both personalized healthcare and population health management. Examples of such work include the prediction of hospital readmission, the identification of adverse effects in high-risk patients, and the creation of personalized disease risk predictions [1]–[5]. These methods rely on the structured clinical data such as disease diagnoses, lab tests, microbiology results, etc. Additionally the data may also incorporate clinical elements such as the medications delivered, and the diagnosis and procedure codes generated as a result of patient care.

While these may appear to be a fairly complete set of medical features there remains a substantial portion of the EMR that, until recently, had been left relatively untapped. This data resides within the text of clinical notes. Clinical notes are documents written by the doctors, nurses and staff providing care to a patient, and offer increased detail beyond what may be inferred from a patient's diagnosis codes. These notes are generated during the standard course of care, and document features such as of the progress of a patient's condition, the plan of care, medical and family history, as well as a number of other clinical attributes. As clinical notes offer such a rich set of ancillary data it is unsurprising that they have drawn the attention of the healthcare informatics community. There are in fact multiple sources of clinical text aggregated as notes within an EMR, ranging from consultations with specialists, to admission details, progress notes, discharge summaries, etc.

Researchers have made significant progress in the extraction of clinically relevant information from these notes. However as noted by Friedman, "most of the systems have been developed specifically for specialized applications and for limited domains [6]". Recent work has attempted to expand the scope of these techniques through the utilization of linguistic tools such as improved lexicon, and complex grammars. However foundational work done by Harris has already established the existence of what are known as sublanguages: *"specialized domains that exhibit specialized constraints due to limitations of the words and relations of the subject matter"* [7].

The assumption that all text extracted from the EMR can be consumed and analyzed in the same manner, regardless of its source, is limiting. The NLP techniques, on which these multi-source systems are based, process data in a statistical manner, thus their ability to produce reliable output is highly dependent on the underlying data. It then stands to reason: if the sources of clinical text are in some way fundamentally different, no high-level linguistic tool will provide an accurate or effective model. This concept is further supported by a more recent work of Friedman, in which they present an analysis of the implications of such specialized domain sublanguages on the development of natural language processing systems [8].

Finally building on the concept of sublanguages, prior work by Stetson et. al [9], provides one of the first quantitative evaluations of structural differences between text of various clinical corpora. Through this work we aim to take the analysis one step further, providing what we believe is the first

comprehensive analysis of the differences in clinical text with relation to the structural, linguistic, and topical features utilized by current NLP models. We will focus on the text of nursing notes, physician notes, radiology and electrocardiogram (ECG) procedure reports. These notes represent the four most prevalent categories of clinical text, and comprise the majority of the patient narrative during an inpatient stay.

Contributions: We provide a brief description of document metrics between each note category including document length and vocabulary differences. Further, we establish the imbalance between the categories of clinical text present in the EMR for an average patient. Next we provide insight into the linguistic differences between each category. We focus on a central set of linguistic features typically utilized to differentiate language models, including part of speech distribution, vocabulary perplexity, and document similarity. Finally, utilizing topic modeling techniques we will demonstrate the underlying themes within each note category are highly distinct to such an extent that they can be utilized to accurately predict the source of previously unseen text.

II. Related work

The desire to structure clinical text has existed long before the emergence of healthcare informatics. Initial work was driven primarily by applications in quality assurance, medical coding, and information retrieval. The ability to automatically provide a consistent set of medical codes across large sets of records was an enticing goal, allowing for both the standardization of codes been institutions and for consistency between patients [10], [11]. Further the emergence of large medical record repositories allowed for increasingly complex research questions. However, to answer these questions required the ability to exact relevant patient records from the repositories. Researchers discovered that to accomplish this task, without the need for manual review of each record, required more detailed information than was currently available in the records' coded sections.

Work by Aronow et al. was one of the first to address these issues directly, noting "The coded portions are used extensively. However, the text portion of the AMRS resource is inaccessible and virtually ignored [12]." Although successful, this and other early work relied heavily on machine learning and statistical inference techniques, such as k-nearest neighbors and tree-based lexical structures to provide structure and basic extraction techniques for freeform clinical text. Recent developments in the field of natural language processing (NLP) have allowed researchers to extend early structuring and extraction methodologies to allow for the extraction of data augmented with clinically accurate contextual information. While new methods are constantly under development, some of the most prominent techniques include boundary detection, part of speech tagging, shallow parsing, entity recognition, morphological reduction, and synonym substitution [13]–[17]. Further, a variety of preprocessing techniques are typically employed to address the noise associated with the natural language aspects of clinical text. These techniques include inverse document frequencies, stemming, stop word removal, and the inclusion of domain-expert annotated "meta-features".

Prior work has noted that although "Machine learning techniques have demonstrated remarkable results in the general domain and hold promise for clinical information extraction, they require large, annotated corpora for training, which are both expensive and time-consuming to generate [18]." As a result, research has focused on the utilization of unsupervised or semi-supervised methods such as topic modeling in order generate large-scale datasets without the need for professional annotation [19]–[21]. It should be noted that a recent set of work has focused on the extraction of features through ontology-based models, that extract clinical data from unstructured text using expertly defined medical ontologies [22], [23]. The hierarchical nature of these ontologies may help to reduce some of noise caused by coding variance. However, there may be a loss in the granularity of the information extracted.

The field of clinical note mining has expanded from simple feature extraction to the modeling of data for predictive tasks such as prediction of a surgical procedure's outcome, identifying patients who had been diagnosed with cardiac disease, and defining the disease severity level of rheumatoid arthritis for use in clinical trials [24]–[26]. These models combine advancements in NLP feature extraction with machine learning supervised methods such as decision trees, neural networks, and support vector machines, as well as unsupervised methods such as association rules, k-means and hierarchical clustering [27], [28]. Today the field continues to progress, yielding work that pushes beyond simple prediction, into the identification of novel artifacts. Such work includes the identification of diagnostic signals that may indicate heart failure, the recognition of adverse drug effects, and even the discovery of previously unknown comorbidies though the creation of extensive disease-symptom networks [29]–[31].

There is also prior work focused on the identification of differences in language use between genders, social groups and individual personalities [32]–[34]. Further, a set of prior work by Harris has provided strong support for the notion that differences in language can be mathematically represented and statistically quantified [7], [35]. Finally, there exists a set of work that focuses on the utilization of linguistic differences to identify various health focused outcomes. These include the *prediction of postpartum changes in emotion*, and *identification of depression in college students* [36], [37]. However, while this work focuses on linguistic differences as a feature to perform their analysis, we aim to employ these differences to produce more accurate processing techniques for healthcare analytics.

III. Data

The data utilized in this work was drawn from the MIMIC III (Multiparameter Intelligent Monitoring in Intensive Care) database, which represents one of the largest sources of publicly available electronic medical records [38]. The database contains approximately 45,000 patient records collected between 2001 and 2012 in conjunction with Boston's Beth Israel Deaconess Medical Center, a 620-bed tertiary academic medical center and a level I trauma center with 77 critical care beds. Patient records were collected across multiple intensive care units (ICU) including medical, surgical, coronary, cardiovascular, trauma, and neonatal. Each record contains an extensive set of clinical features including patient's physiologic signals, chart data, vital signs, as well as time series data captured from a patients' bedside monitor. Further, the records

provide the set of clinical notes recorded for a patient over their stay. Between all patients, the database contains just over 2.4 million notes.

Each note provides an annotation denoting the specific category of clinical text it represents, such as a nursing note or a radiology report. For our analysis we utilize the most prominent note categories, defined as those representing over 5% of the total note instances. This criteria provides four distinct categories of notes: nursing notes, physician notes, radiology and ECG reports. These categories represent two major classes of text, clinical text *(nursing, physician)* and procedural reports *(radiology, ECG)*. Together these categories provide an highly comprehensive view of clinical text, accounting for just over 93% of all notes in the database, with 1,046,053, 141,624, 870,504 and 209,051 instances in the nursing, physician, radiology and ECG categories respectively.

A. Preprocessing

Although the MIMIC database is well maintained, as with most natural language tasks, the notes required a set of preprocessing tasks to obtain text fit for analysis. First, in compliance with standards set forth by the Health Insurance Portability and Accountability Act the names of all patients, doctors and nurses were deidentified, as were all dates. As an example, the strings *[**3069-3-16**]*, and *[**Known patient firstname **] [**Last Name (NamePattern4) 1716**]* represent a deidentified date and patient name respectively. In order to prevent the text and structure of these redacted elements from skewing our analyses they were removed though the use of regular expression parsing.

Next, the text was stripped of all digits. The proper handling of numerical elements in text remains an open question in the NLP community. There is debate as to the numerics should be removed entirely, or replaced with a constant placeholder. However in a medical context numerics can represent a multitude of different entities, including dosages, weights, counts, times, frequencies, or rates. As such, utilizing a common placeholder would replace the numeric segments of *1*-mg tablet, a *85* kg patient, and a medication administered at *22:00* with the same linguistic element. This transformation has the potential to incorrectly bias the interpretation of analyses, particularly those which utilize vocabulary similarities or rely on normalized word frequencies.

Additionally, it is well established that clinical text often presents varying levels of fragmentation and grammatical correctness. As such, to ensure that our analyses reflect the variations between note categories and not the clinicians' documentation styles, all punctuation was subsequently removed. It should be noted that all tokenized text was converted to lowercase lettering in an effort standardize string comparisons. Finally a set of stopwords (high frequency, low information words) were removed from each note. The stopword list was comprised of standard English stopwords, augmented with a medical stopword list obtained from NCBI PubMed [39].

IV. Methods

This work focuses on three distinct evaluations of clinical text. We begin with an evaluation of the structural differences among each of the four note categories. Next we evaluate the differences between each note category using common NLP metrics. Finally, we investigate how the underling topics vary within each category. The methodology for each evaluation can be found in the respective sections below, while the results and discussion can be found in the corresponding elements of sections V and VI.

A. Note Structure

We began our investigation with an evaluation of document length, focusing on the average word count of each category. The text of each note was tokenized using the Natural Language Toolkit (NLTK) python package [40]. The average document length was then computed, and an unpaired *t*-test performed between all combinations of note categories. As clinical notes provide detailed information of a patient encounter, it stands to reason the length of a note is highly associated with the severity and complexity of a patient's condition or treatment. To account for this variability we perform an additional analysis, removing those notes with an outlier number of words. As noted prior, clinical text typically lacks correct or consistent punctuation. As a result, we found that while a sentence-level evaluation was technically feasible, the results would yield meaningless values for interpretation, and as such were not included in this work.

Outliers were identified using the median absolute deviation (MAD). MAD is a highly robust metric of variability, similar to standard deviation. However, unlike standard deviation, MAD is based on the median value. This distinction acts to reduce the effect of extreme outliers, which is particularly important in heavy tailed distributions where the highest values may be orders of magnitude larger than the median. The utilization of robust methods, such as MAD, over the normal standard deviation has been explored in detail by Iglewicz et. al [41]. For all analyses in this work, outliers are selected at a threshold of ± 3 MAD.

Noting significant differences in their document lengths, we then moved to investigate the vocabulary for each of the note categories. To accomplish this task we first identified the total word count and set of unique words for each category. Next, we computed the symmetric difference between the set of unique words for each category pair. The symmetric difference provides the number of terms which appear in one category or the other, but not in both. This value was then normalized with respect to the total size of both vocabularies, providing a metric for the proportion of overlap in terms utilized between two categories. With this metric, a value of 1 would represent two completely distinct vocabularies, where a value of 0 would indicate the two vocabularies were identical.

Although this work intends to highlight fundamental differences among the note categories themselves, it is also important to examine the distribution of these categories within a patient's set of clinical text. An awareness of a potential category imbalance would be critical for future work, which may utilize differing processing or weighting methods between each of the categories. To address this question we calculated the proportion of each note category for each patient, averaging across all patients. However it is important to recognize that not all patients may have clinical text from each category, particularly within the procedural report class. As such we also

calculate the average count of each category ignoring patients who have no documented notes in a particular category. This framework shifts the result interpretation slightly. As an example for radiology reports, the initial analysis would present the average proportion of radiology reports per patient, where as the second would express the average category proportion for those patients who have had at least one exam.

B. Linguistic Features

As our prior analyses were able to quantitatively demonstrate notable differences between the vocabulary of each category, we then moved to investigate the possibility of linguistic differences between the categories. We first analyzed the distribution of the parts of speech used for each word across the different note categories. Part of speech tags have been previously established as an important aspect of many NLP applications‘including “syntactic parsing, named entity detection, and other information extraction tasks" [42]. Identifying deviations between the part of speech distributions between different note categories would further support the notion that additional consideration must be given to the source of clinical data in order to provide accurate contextual analysis. The part of speech tags were determined using the NLTK tagger, which provides labels from the widely employed Penn Treebank tagset. The comparison between note categories was performed using the Chi-Squared goodness of fit test, with all low frequency tags (under an expected value of 5) pooled into an *other* category.

Next to provide more formalism to the vocabulary analysis found in the *Note Structure* section, we compared the vocabularies of the clinical (*Nursing and Physician*) and procedural (*Radiology and ECG*) note classes using the cosine similarity, a widely utilized NLP document similarity metric. To calculate the similarity the note text was put through a *term frequency–inverse document frequency (tf-idf)* transformation, which provides a document vector where each term in the note text is inversely weighted to its observed global frequency within the category. From here the cosine similarity was computed between each categories' document matrix. Additionally we took the evaluation one step further, breaking the comparison down from a class level to evaluate the similarity between each of the four individual note categories.

We then explored one of the most prominent NLP metrics for quantifying the complexity of a vocabulary: its perplexity. In the context of NLP, perplexity provides a metric for the probability of predicting a term that occurs in a corpus. By this definition simpler corpus lend themselves to lower perplexities, as the probability of correctly predicting a future term is higher due to the limited vocabulary size. Perplexity is an important metric with respect to multi-sourced text, as it provides a measure of a models ability to represent the vocabulary. Thus to accurately represent two sources with highly different perplexities may require substantially different models, providing the ability to capture the complexities of each without the risk of over- or under-fitting the respective vocabularies. It should be noted perplexity can be measured with respect to a n-gram language model where the probability of correctly predicting the next term is based on the n-1 terms before it. The average perplexity for each class was computed independently using 10-fold cross validation, utilizing Good Turing smoothing. Good Turing smoothing was chosen as to avoid the bias of determining a smoothing factor required by other techniques such as Laplace (Additive) smoothing. Further, as the content of a clinical note is highly complex it may be insufficient to evaluate the complexity of the vocabularies based on a unigram model. To account of this, the analysis was extended to the bigram models of each category. Again, statistical significance between each category's perplexity was determined using an unpaired *t*-test.

Finally, it is well established that real-world natural language vocabularies are inherently noisy, containing many low frequency words and phrases. In an effort to reduce this noise, many NLP applications utilize a subset of the highest frequency terms across the vocabulary's they model. To evaluate the effect of this preprocessing technique on differences previously established by this work, we calculated the similarity across subsets of the *Top-N* terms for each category. While previously the vocabulary similarity was quantified by cosine similarly, in this analysis the similarities were computed utilizing the Jaccard coefficient. This interchange was strategic, as the goal of this analysis was to highlight the effects of the *Top-N* preprocessing technique on eliminating low frequency terms between each category, as opposed to the distribution of terms between each category as was required in the prior analysis. While the cosine similarity accounts for the global term frequency within all notes of a specific category, the Jaccard coefficient calculates only the overlap between the term sets. The Jaccard coefficient can be defined as $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$ [43], where sets A and B represent the unique set of terms found in the *Top-N* terms in each category subset, and the value of N was varied between the top 5 to 1,000,000 terms.

C. Topic Modeling

Although it is straightforward to determine topics of a single note category, it is particularly difficult to quantify differences between topics discovered from multiple categories through a direct comparison. Thus to determine a metric for the differences in the underlying topics discovered from each note category we transformed the problem into a classification task. To transform these topic models into a classification task the training notes were first aggregated into four separate documents, one containing the text of all notes in each category. Testing documents were left as individual notes to prevent biasing the result by providing the model additional terms during prediction. Next we trained a topic model over the four documents. Using this model we then computed the similarity of each test note to each of the four training documents. Then, acting as a classification task, the category of each testing note was assigned to one of the four training documents based on the highest degree of similarity. The resulting accuracy scores then act as a proxy for the the degree of separability between the topics of each category. Highly overlapping topic spaces would produce similar cosine similarity score between the training documents (note categories), increasingly the likelihood of a misclassification, where as highly distinct underlying topics would have significantly more polarized similarity scores within the topic-space, and thus produce improved classification results. .

To generate the topics we utilize a model known as

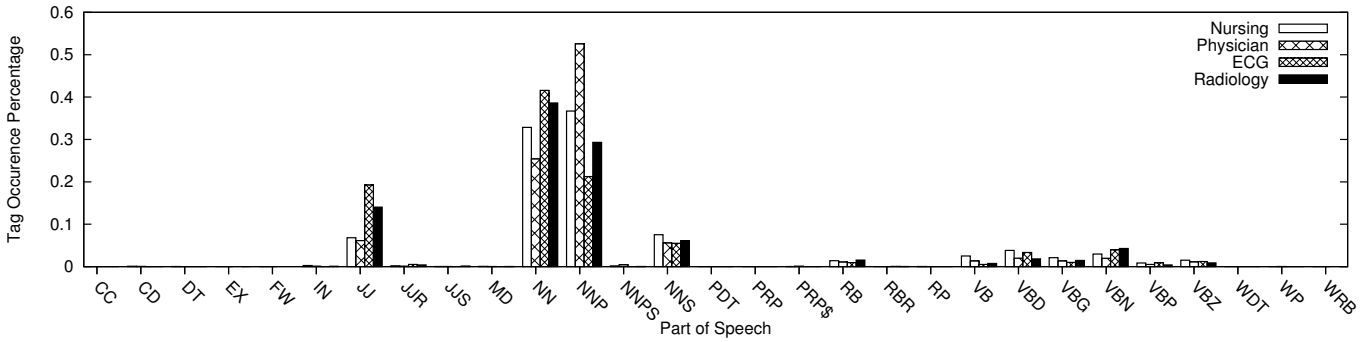

Fig. 1: Part of Speech Distribution by Category

Note Type	Mean Word Count	Word Count SD	Total Notes
Nursing	94.98 *(93.55)*	76.20 *(74.94)*	1,046,001 *(1,019,606)*
Physician	480.92 *(467.24)*	182.04 *(158.18)*	141,624 *(138,280)*
Radiology	150.27 *(107.68)*	120.22 *(55.01)*	870,504 *(733,664)*
ECG	18.60 *(18.58)*	9.29 *(9.27)*	209,010 *(205,331)*

TABLE I: Average Document Length

	Normalized Symmetric Difference (%)					
	Nursing	**Physician**	**Radiology**	**ECG**	**Total Words**	**Total Unique Words**
Nursing	0	81.84	86.50	98.65	99,353,821	696,934
Physician	81.84	0	71.80	94.56	68,110,313	143,237
Radiology	86.50	71.80	0	94.89	130,813,301	161,724
ECG	98.65	94.56	94.89	0	3,888,290	8,373

TABLE II: Vocabulary Differences by Category

Latent Semantic Indexing (LSI) provided by the python topic modeling package GenSim [44]. The LSI model was chosen as it not only allows us to extract the underlying topics, but also to directly compute a cosine similarity of a new document with the indexed training documents. As such, we utilized the the argmax of the similarities to assign the note a category label. The analysis was performed using stratified 10-fold cross validation, where the LSI-training documents were regenerated at each fold, and as is standard within most document retrieval applications performance is measured using the precision and recall.

V. Results

The results of each analysis detailed in section IV can be found in the respective sections below.

A. Note Structure

The results of the document length evaluation can be found in Table I. The average word count and standard deviation are reported for both the complete set of notes and for notes excluding those with an outlier number of words *(denoted by the italicized term within parentheses)*. Differences in document length were found to be statistically significant between all pairs of note categories at 95% confidence. It should be noted that category comparisons were not performed between averages containing outliers and those with them removed.

Next, Table II provides the results for the vocabulary analysis. Within this table the first four columns present the normalized symmetric differences between the unique terms found in each category. As noted prior, the symmetric difference is the set of terms existing in either category, but not both, and is normalized by the total unique word count of both categories. The final two columns represent the count of total words, and unique words present in each category respectively.

Finally, the average category distribution between patients' clinical text can be found in Table III. Each of the note categories is partitioned into two rows, representing each of the analyses. The upper row provides the average note count statistics across all patients, while the bottom line provides the statistics for only those patients who have at least one note for the respective category.

Note Type	Mean Category %	SD	Total Notes
Nursing	41.70	33.25	58,006
	55.13	26.85	43,872
Physician	04.11	10.70	58,006
	26.55	11.99	8,983
Radiology	37.08	31.32	58,006
	47.25	27.74	45,526
ECG	17.10	22.86	58,006
	22.45	23.78	44,185

TABLE III: Average Category Proportion

B. Linguistic Features

We began the linguistic analysis with an investigation into the differences between each categories' parts of speech distribution. Figure 1 represents the normalized tag proportion for each category. The tags were drawn from the Penn Treebank Tag Set, and are presented in alphabetical order.

To extend the analysis of vocabulary similarity we calculated the cosine similarity between the two major classes of note text, clinical and procedural, finding a similarity of 0.283. With such a substantial distance between the tf-idf document vectors, we repeated the analysis breaking the vectors down into the individual categories. The similarity values between all-pairs of note categories can be found in Table IV. Finally we broke the analysis down further, calculating the similarity between individual notes, stratified to preserve the category proportions. A visualization of the similarity matrix can be seen in Figure 2, where higher similarity values are represented by brighter colors. Sections A, B, C, and D represent notes

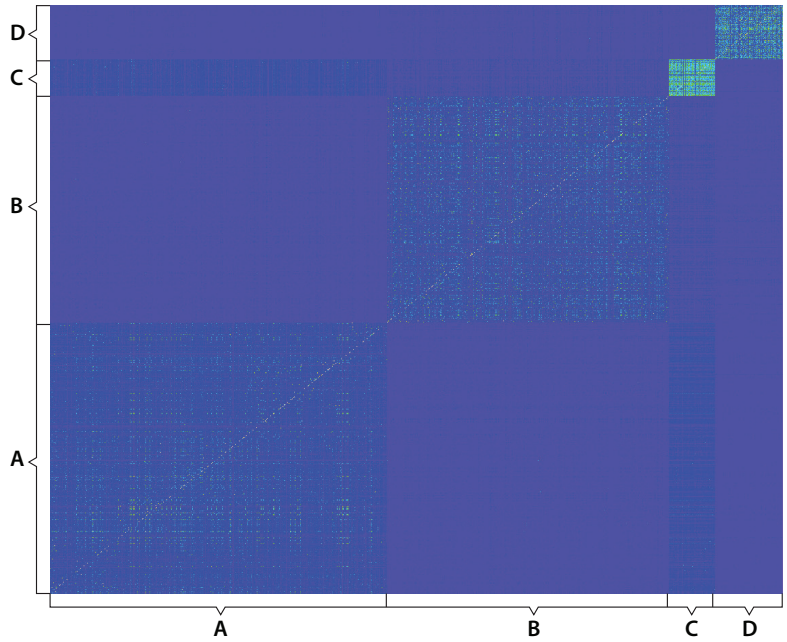

Fig. 2: Individual Note Cosine Similarity. Section A: Nursing Notes, B: Radiology Reports, C: Physician Notes, D: ECG Reports

from the categories of Nursing, Radiology, Physician, and ECG respectively.

Note Type	Nursing	Physician	Radiology	ECG
Nursing	1	0.4697	0.2070	0.0500
Physician	0.4697	1	0.2723	0.0877
Radiology	0.2070	0.2723	1	0.1160
ECG	0.0500	0.0877	0.1160	1

TABLE IV: Category Vocabulary Cosine Similarity

Next the average perplexity for each note category can be found in Table V. Each category is partitioned into two rows, where the upper row provides the average unigram perplexity, while the lower line provides average bigram perplexity.

Note Type	Mean Perplexity	Perplexity SD
Nursing	1,411.40	105.09
	107.33	10.54
Physician	1,324.72	33.78
	29.23	1.25
Radiology	762.31	5.85
	35.68	0.03
ECG	164.26	3.36
	8.60	0.16

TABLE V: Average Category Perplexity

The final linguistic analysis was focused on the noise reduction technique of utilizing the only the *Top-N* words in each category. To quantify how this technique would coalesce the note categories vocabularies we calculated the Jaccard Similarity between the set of *Top-N* terms, varying N from 5 to 1,000,000. Figure 3 presents the results of this analysis on log-scale in order to smooth the visualization.

C. Topic Modeling

Table VI presents the average precision, recall and F_1 - Score for the topic prediction analysis. For clarity, a correct prediction represents the following scenario. A test note, stripped of it category, is converted into LSI-space using the trained topic model. The cosine similarity is then computed against each of the four documents, representing each of the four possible note categories. The argmax of the calculated similarity scores correctly aligns with the notes true category.

Additionally, as topic models are often viewed as "black-boxes", we aimed to provide insight into the underlying topics discovered from the note text. Each of the models topics can be found in Table VII, along with the top 5 terms which contribute to the direction of the topic in both a positive and negative aspect, also provided are the terms corresponding weights.

Note Type	Average Precision	Average Recall	Average F1
Nursing	0.9982	0.9724	0.9850
Physician	0.8797	0.9859	0.92701
Radiology	0.9914	0.9999	0.9957
ECG	0.9978	0.9998	0.9988

TABLE VI: Topic Model Category Prediction Performance

Topic Number					
1	pt	left	pm	mgdl	ct
	0.2724	*0.1945*	*0.1907*	*0.1432*	*0.1263*
2	tracing	previous	sinus	rhythm	wave
	0.4777	*0.3135*	*0.3027*	*0.2320*	*0.2282*
3	reason	contrast	pt	tracing	ct
	0.2716	*0.2100*	*-0.2086*	*-0.1941*	*0.1800*
4	pt	mgdl	pm	meql	icu
	-0.3440	*0.2859*	*0.2338*	*0.2251*	*0.1558*

TABLE VII: Most Influential Terms by Topic

VI. Discussion

In the opening chapter of their foundational book on corpus linguistics, Biber et al. note that "corpus-based analyses must

go beyond simple counts of linguistic features. That is, it is essential to include qualitative, function interpretations of quantitative patterns [45]." It is through this lens that we approach the discussion of our analyses.

A. Note Structure

Beginning with the average note length we find that the average word count of a note varies widely, even between notes of the same category. This variation is likely the result of a number of factors, including the severity of the patient's condition, the type of encounter being documented (routine rounds vs. admission to an ICU), and the time since the last documentation. Although all categories were statically different in their average word length, we can dive deeper, focusing on the notes within the longest (physician) and shortest (ECG) categories.

Starting with the shortest category, the brevity of ECG reports is somewhat striking. A review of ECG notes highlights multiple instances of phrases such as "ECG interpreted by ordering physician," suggesting an extended interpretation of the ECG itself would be found in physician note category. While this may account in part for the brevity of the category, we then looked to the second procedural note category for comparison. The workflow of ECG and radiology reports are similar in that they are both written by the reading cardiologist and radiologist respectively, and given to the ordering physician for interpretation. However we find the average document length of radiology reports to be much longer, indicating that the brevity of the ECG may be related to other intrinsic factors. Closer examination of the ECG note text itself reveals text highlighting an extremely specific set of physiological attributes, such as atrial flutter and sinus rhythms. This limited vocabulary is a category feature we will revisit a number of times throughout this discussion.

On the other side, the extended length of physician notes proves to be an interesting characteristic. As noted prior physician notes may contain information beyond direct patient observations, such as the interpretation of procedural reports. Further, examination of these notes reveals that they often contain sections detailing a patient's medical histories, assessment and plan of care, which may account for the increase in word count. Additionally, as will see in the distributional analysis, the presence of a physicians note may itself be indicative of a subgroup of patients.

Moving to the vocabulary analysis we observe that despite the substantial number of words that comprise the vocabulary of each category, the words used within each are remarkably different. The normalized symmetric differences, found in Table II, reveal less than a 30% overlap between even the most similar categories. Looking next to the similarity score between each category pair, we can highlight patterns that may provide insight into the causes of these deviations. Beginning with nursing notes, it is unsurprising that the category with the highest similarity is that of physician notes, as both categories detail patient observations and likely contain similar clinical vocabulary. However this pattern is not reciprocal, as physician notes are in fact most similar to radiology reports. This result may be an artifact of the clinical workflow that generates these reports. As noted prior, procedural reports are generated by a specialist, and provided back to the ordering physician for interpretation. It then stands to reason that physician notes would utilize a high proportion of words from the radiology report in their own notes. Following this logic, we look to the second procedural note category of ECG reports, written by the reading cardiologist, and we find that ECG reports are in fact most similar in vocabulary to physicians notes.

As with the two clinical text categories, we find a high similarity between the ECG and radiology procedural reports. This is to be expected as these reports detail procedures, and likely have much a similar set of base terminology. Despite these similarities, ECG reports are in fact highly different from all other categories, with differences all above 94%. One particularly interesting observation is that of the 5.7 million words found between the ECG reports, we find only around 8,700 unique words, again demonstrating the highly specific language utilized in the notes of this category.

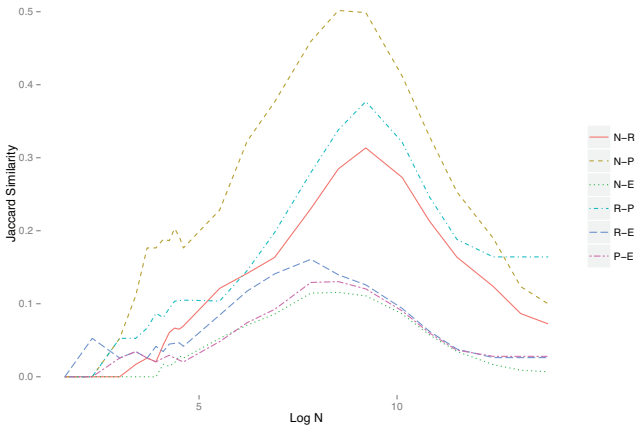

Fig. 3: *Top-N* term Similarity between Categories

Our final structural analysis investigated the average note category distribution for a patient. Although not focused on the clinical text directly, the distribution analysis is especially important for the consideration of future work. Building on the idea that due to the fundamental differences highlighted in this work, differing note categories will require specific processing techniques, we must then also be aware of the imbalance between the categories themselves. Just as imbalanced classes must be handled appropriately in the context of classification tasks, the imbalance within the note categories would then need to be addressed during any NLP preprocessing and model training tasks.

These results highlight an interesting phenomenon, although the average proportion of nursing, radiology and ECG notes demonstrate only a marginal increase compared to the averages for patients containing at least one note in their respective category, physician's notes show a drastic increase of approximately 6.5x. The magnitude of this disparity suggests that patients who have at least one physician encounter, may in fact require multiple physician evaluations over the course of their admission. An awareness of variations such as these will be particularly important in the development of successful NLP models, as it demonstrates that class imbalance may not be a global attribute. As a result, techniques which address

this imbalance may require adjustment on a patient by patient basis. This scenario fosters the idea that a single model may be insufficient, but rather an ensemble of models may be required for appropriate processing.

B. Linguistic Features

We began out investigation into the linguistic features of differing note categories by evaluating the parts of speech used within each category. Within both categories of the clinical text class we find the most prominent part of speech tag represents *proper noun, singular (NNP)*, followed by *noun singular or mass (NN)*. However, within the categories of the procedural report class this pattern is reversed, with the common noun *NN* as the leading tag, followed by the proper noun *NNP*. This distinction is particularly important for applications such as named entity recognition and disambiguation, which have shown great promise in recent applications to clinical text information extraction. However, prior work has established the importance of processing proper and common nouns to achieve optimal performance of such methods, particularly for situations where proper nouns may contain direct names but common nouns employ phrases such as "the patient" to refer to the same entity [46].

Further we note two tags that are notably over-represented in the procedural report class. These include the adjective tag (JJ), which comprises 19.25% and 14.03% of tags in the ECG and radiology categories, when compared to the 6.80% and 6.12% in the nursing and physician categories respectively. Additionally the tag VBN (verb past participle) averages 4.11% of tags between the procedural report categories, compared to an average of 2.34% between the two clinical text categories. While these differences may seen trivial the increase in adjectives is an important feature in a clinical context. For example the presence of a tumor, can be quantified by the adjectives of *benign* or *malignant*, indicating two very different conditions.

Expanding on the vocabulary differences, we then quantified the document similarity between the clinical and procedural classes. As would be expected, text coming from a clinical class is highly different from that found in the categories comprising the procedural reports. However, breaking down the classes into the four categories reveals that even within a class the note categories are substantially different, with similarities of 0.4697 and 0.1160 between Nursing/Physician notes and ECG/Radiology reports respectively. One particularly interesting aspect of the analysis was the granularity at which these differences are notable. As established by the prior analyses the overall vocabulary of the categories is remarkably different on a macro level. However, we have not yet investigated the categories on a note-by-note basis. Examining the similarity matrix found in Figure 2 we find a clear distinction between the documents of different classes, demonstrating that even with only a few hundred words the categories are clearly distinguishable. It should be noted that a visible band exists between sections A and C, representing similarity between nurse and physician notes. This elevated similarity is expected, as multiple notes may be detailing the same patient condition from both the nurse and physician perspective. However overall these inter-category similarities represent lower similarity scores than the intra-category comparison, as denoted by the darker shades of blue.

Next, the perplexity analysis allows us to further quantify each category's language complexity. At a high level we note a decrease in perplexity from unigram to bigram models. This is a property of perplexity, as the probability of correctly predicting a word is helped substantially by knowledge of the prior word, particularly in complex vocabularies, such as those detailing patient conditions and hospital procedure reports. One interesting observation is the change in perplexity between the unigram and bigram models of the physician notes and radiology reports. In a unigram model physician notes present a significantly higher perplexity than radiology reports. The increased difficulty in predicting a single word is supported by physician notes increased proportion of unique-words to total-words detailed in Table II in comparison to the radiology reports. However when moving to a bigram language model this pattern is reversed, with radiology reports exhibiting higher perplexity than physician notes.

This distinction is particularly important for NLP models that extend beyond bigrams to the concept of *n-grams* in order to increase vocabulary information during training. Differing perplexity patterns such as these may indicate the need to utilize NLP models which vary the n-gram level based on the category of clinical text being processed in order to achieve optimal performance.

Finally an evaluation of the *Top-N* word similarity between the categories allows us to extend the vocabulary analysis to examine how the established differences vary after undergoing a standard noise reduction technique. Looking at the results in Figure 3 it is clear that although the subset of vocabularies can help coalesce text from different categories, they do remain overall remarkably different, with no value of *N* exceeding a similarity of 50%. Beginning with the highest frequency word-sets (lowest n-values) in each category we find the sets to be extremely disparate, with similarities below 23% for set including up to the top 250 words per category. Increasing the size of the word-sets further does shows a marked increase in category similarities. This is likely as result of these expanded containing universal medical terminology that is common to all categories. Finally continuing to increase the size of the word-sets to approach the full vocabulary size we note a decrease in similarity. The sharp decrease is likely caused by the inclusion of low frequency words, and of terminology specific to the respective category. We find this result to be particularly noteworthy as it demonstrates even with the noise reduction techniques of stop word removal and word frequency filtering the language used in each category is still extremely different, supporting the notation that additional processing techniques are required for proper language modeling of the heterogeneous sets of clinical text categories.

C. Topic Modeling

In our final analysis we examine the underlying topics found within each note category. As the comparison of topic models remains an open area of research, we created an experimental design in which the predictability of notes projected into LSI space could be utilized as a proxy for the degree of separability between the topics of each category. The results

demonstrate extremely high precision and recall, indicating the discovered topics were remarkably distinct. It should be noted that the lowest predictability is found within the physician category. Manual inspection of the confusion matrices generated at each fold reveals, that as would be expected, the most common incorrect class assignment for a physician notes is that of a nursing note. This misprediction is logical as both note categories detail aspects of the patients condition, and their terminology similarity has been demonstrated in both vocabulary analyses.

Next we expand the topic model analysis moving to examine the specific terms which influence the category for each discovered topic, noting some clear patterns. Looking to Table VII the terms within topic 2 appear to denote terminology commonly found in the ECG category, highlighting words such as 'sinus, rhythm, and wave'. For clarity, the following is an example of an actual ECG note: *"Sinus rhythm. Right bundle-branch block. Left anterior fascicular block. Compared to the previous tracing of sinus rhythm has appeared. TRACING #2"*

Further we note some interesting differences in terminology between topic what may have a relation to the vocabularies of different note categories. Topic 1 is most influenced (positively, weight 0.272) by the appearance of the term *pt*, whereas topic 4 is most influenced (negatively, weight -0.344) by the appearance of the same term. For clarity a negative influence decreases the likelihood that a note will be assigned to a particular topic, and in the classification framework provides a lower cosine similarity which decreases the probability the note will be assigned to the respective category.

Reviewing the most frequent terms from each class we find the term *pt* is the most common term in the nursing category, while in the physician category it represents the 11^{th} most common term. A closer examination reveals the 6^{th} most common term within the physician category to be the full word *patient*. Moving further, we find the proportion of occurrence for the terms *pt* and *patient* are roughly equal within physician notes, constituting around 0.5% of all terms in the category. However the interchangeability of these terms does not extend to the nursing category. Within the text of nursing notes the term *pt* constitutes just over 2% of all terms, whereas the full term *patient* comprises a minimal 0.3%.

The ability to capture nuanced language and notation differences such as these are critical for effective processing of clinical text. While it may be possible to coalesce shorthand such as *pt* to the full term *patient* through the use of manually curated lists, this would require additional preprocessing not performed by the majority of existing work. Further, it would be naive to believe that a complete list of transformations could be maintained for text from multiple categories.

VII. Conclusion

Reflecting back, this work provides a comprehensive evaluation of the structural, linguistic, and topical features of four prominent clinical text categories. The analyses provided by our work have demonstrated fundamental differences in each of the evaluation areas, across all note categories. A deep understanding of clinical text is particularly important in the context of personalized care, as prior work has demonstrated that no coding system is currently sufficient to allow clinicians to describe even a single diagnosis completely and accurately [47], [48]. As a result an awareness of the differences in types of clinical text will become increasingly important as natural language becomes increasingly intertwined with the field of healthcare informatics.

Unfortunately acknowledgement of these differences is not yet enough, a significant effort remains to truly understand how clinical text and NLP can best be utilized in conjunction with informatics methodologies to provide contextually relevant and accurate analysis of patient conditions. However, it is our hope that this work inspires others to think about the data utilized in their analyses, and provides a foundation for which future work can build to continue the advancement of patient care.

Acknowledgements

This work is supported in part by the National Science Foundation (NSF) Grant IIS-1447795

References

[1] D. He, S. C. Mathews, A. N. Kalloo, and S. Hutfless, "Mining high-dimensional administrative claims data to predict early hospital readmissions," *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 272–279, 2014.

[2] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, "Big data in health care: using analytics to identify and manage high-risk and high-cost patients," *Health Affairs*, vol. 33, no. 7, pp. 1123–1131, 2014.

[3] D. A. Davis, N. V. Chawla, N. A. Christakis, and A.-L. Barabási, "Time to care: a collaborative engine for practical disease prediction," *Data Mining and Knowledge Discovery*, vol. 20, no. 3, pp. 388–415, 2010.

[4] T. H. McCormick, C. Rudin, D. Madigan *et al.*, "Bayesian hierarchical rule modeling for predicting medical conditions," *The Annals of Applied Statistics*, vol. 6, no. 2, pp. 652–668, 2012.

[5] N. V. Chawla and D. A. Davis, "Bringing big data to personalized healthcare: a patient-centered framework," *Journal of general internal medicine*, vol. 28, no. 3, pp. 660–665, 2013.

[6] C. Friedman, "A broad-coverage natural language processing system." in *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2000, p. 270.

[7] Z. Harris, "Theory of language and information: a mathematical approach," 1991.

[8] C. Friedman, P. Kra, and A. Rzhetsky, "Two biomedical sublanguages: a description based on the theories of zellig harris," *Journal of biomedical informatics*, vol. 35, no. 4, pp. 222–235, 2002.

[9] P. D. Stetson, S. B. Johnson, M. Scotch, and G. Hripcsak, "The sublanguage of cross-coverage." in *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2002, p. 742.

[10] N. Sager, M. Lyman, C. Bucknall, N. Nhan, and L. J. Tick, "Natural language processing and the representation of clinical data," *Journal of the American Medical Informatics Association*, vol. 1, no. 2, pp. 142–160, 1994.

[11] L. S. Larkey and W. B. Croft, "Automatic assignment of icd9 codes to discharge summaries," *University of Massachusetts*, 1995.

[12] D. Aronow, S. Soderland, J. Ponte, F. Feng, W. Croft, and W. Lehnert, "Automated classification of encounter notes in a computer based medical record." *Medinfo. MEDINFO*, vol. 8, pp. 8–12, 1994.

[13] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, 2010.

[14] R. A. Erhardt, R. Schneider, and C. Blaschke, "Status of text-mining techniques applied to biomedical text," *Drug discovery today*, vol. 11, no. 7, pp. 315–325, 2006.

[15] D. Heinze, M. Morsch, R. Sheffer *et al.*, "A natural language processing system for medical coding and data mining," in *AAAI-Twelfth Innovative Applications of Artificial Intelligence Conference*, 2000.

[16] X. Zhou, H. Han, I. Chankai, A. Prestrud, and A. Brooks, "Approaches to text mining for clinical medical records," in *Proceedings of the 2006 ACM symposium on Applied computing*. ACM, 2006, pp. 235–239.

[17] G. Hripcsak, S. Bakken, P. D. Stetson, and V. L. Patel, "Mining complex clinical data for patient safety research: a framework for event discovery," *Journal of biomedical informatics*, vol. 36, no. 1, pp. 120–130, 2003.

[18] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, J. F. Hurdle *et al.*, "Extracting information from textual documents in the electronic health record: a review of recent research," *Yearb Med Inform*, vol. 35, pp. 128–44, 2008.

[19] A. Salleb-Aouissi, A. Radeva, R. Passonneau, A. Tomar, D. Waltz *et al.*, "Diving into a large corpus of pediatric notes," *Proc. ICMLWorkshop on Learning from Unstructured Clinical Text*, 2011.

[20] R. Cohen, I. Aviram, M. Elhadad, and N. Elhadad, "Redundancy-aware topic modeling for patient record notes," *PloS one*, vol. 9, no. 2, p. e87555, 2014.

[21] Z. Wang, A. D. Shah, A. R. Tate, S. Denaxas, J. Shawe-Taylor, and H. Hemingway, "Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning," *PLoS One*, vol. 7, no. 1, p. e30412, 2012.

[22] C. Tao, G. Jiang, T. A. Oniki, R. R. Freimuth, Q. Zhu, D. Sharma, J. Pathak, S. M. Huff, and C. G. Chute, "A semantic-web oriented representation of the clinical element model for secondary use of electronic health records data," *Journal of the American Medical Informatics Association*, vol. 20, no. 3, pp. 554–562, 2013.

[23] W. Hsu, R. K. Taira, S. El-Saden, H. Kangarloo, and A. A. Bui, "Context-based electronic health record: toward patient specific healthcare," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 16, no. 2, pp. 228–234, 2012.

[24] R. Cobb, S. Puri, D. Z. Wang, T. Baslanti, and A. Bihorac, "Knowledge extraction and outcome prediction using medical notes," in *ICML workshop on Role of Machine Learning in Transforming Healthcare*, 2013.

[25] T. E. Perry, H. Zha, K. Zhou, P. Frias, D. Zeng, and M. Braunstein, "Supervised embedding of textual predictors with applications in clinical diagnostics for pediatric cardiology," *Journal of the American Medical Informatics Association*, vol. 21, no. e1, pp. e136–e142, 2014.

[26] C. Lin, E. W. Karlson, H. Canhao, T. A. Miller, D. Dligach, P. J. Chen, R. N. G. Perez, Y. Shen, M. E. Weinblatt, N. A. Shadick *et al.*, "Automatic prediction of rheumatoid arthritis disease activity from the electronic medical records," *PloS one*, vol. 8, no. 8, p. e69932, 2013.

[27] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," *Nature Reviews Genetics*, vol. 13, no. 6, pp. 395–405, 2012.

[28] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, and L. Hua, "Data mining in healthcare and biomedicine: a survey of the literature," *Journal of medical systems*, vol. 36, no. 4, pp. 2431–2448, 2012.

[29] R. J. Byrd, S. R. Steinhubl, J. Sun, S. Ebadollahi, and W. F. Stewart, "Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records," *International journal of medical informatics*, vol. 83, no. 12, pp. 983–992, 2014.

[30] S. V. Iyer, R. Harpaz, P. LePendu, A. Bauer-Mehren, and N. H. Shah, "Mining clinical text for signals of adverse drug-drug interactions," *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 353–362, 2014.

[31] Y. Ling, Y. An, and X. Hu, "A matching framework for modeling symptom and medication relationships from clinical notes," in *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*. IEEE, 2014, pp. 515–520.

[32] M. L. Newman, C. J. Groom, L. D. Handelman, and J. W. Pennebaker, "Gender differences in language use: An analysis of 14,000 text samples," *Discourse Processes*, vol. 45, no. 3, pp. 211–236, 2008.

[33] E. Cunha, G. Magno, M. A. Gonçalves, C. Cambraia, and V. Almeida, "A linguistic characterization of google+ posts across different social groups."

[34] J. W. Pennebaker and L. A. King, "Linguistic styles: language use as an individual difference." *Journal of personality and social psychology*, vol. 77, no. 6, p. 1296, 1999.

[35] Z. S. Harris, *A grammar of English on mathematical principles*. John Wiley & Sons Inc, 1982.

[36] S. Rude, E.-M. Gortner, and J. Pennebaker, "Language use of depressed and depression-vulnerable college students," *Cognition & Emotion*, vol. 18, no. 8, pp. 1121–1133, 2004.

[37] M. De Choudhury, S. Counts, and E. Horvitz, "Predicting postpartum changes in emotion and behavior via social media," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013, pp. 3267–3276.

[38] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark, "Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database," *Critical care medicine*, vol. 39, no. 5, p. 952, 2011.

[39] PubMed Help [Internet]. Bethesda (MD). (2005) [table, stopwords] available. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T43/

[40] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python*. " O'Reilly Media, Inc.", 2009.

[41] B. Iglewicz and D. C. Hoaglin, *How to detect and handle outliers*. Asq Press, 1993, vol. 16.

[42] D. Jurafsky and J. H. Martin, *Speech and language processing*. Pearson, 2014.

[43] G. Salton and M. J. McGill, "Introduction to modern information retrieval," 1986.

[44] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.

[45] D. Biber, S. Conrad, and R. Reppen, *Corpus linguistics: Investigating language structure and use*. Cambridge University Press, 1998.

[46] S. Zhang and N. Elhadad, "Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts," *Journal of biomedical informatics*, vol. 46, no. 6, pp. 1088–1098, 2013.

[47] W. R. Hogan and M. M. Wagner, "Free-text fields change the meaning of coded data." in *Proceedings of the AMIA Annual Fall Symposium*. American Medical Informatics Association, 1996, p. 517.

[48] H. Quan, B. Li, L. Duncan Saunders, G. A. Parsons, C. I. Nilsson, A. Alibhai, and W. A. Ghali, "Assessing validity of icd-9-cm and icd-10 administrative data in recording clinical conditions in a unique dually coded database," *Health services research*, vol. 43, no. 4, pp. 1424–1441, 2008.