Exploring insights and tagging medical narratives using unsupervised learning

Siddharth Rayabharam
Masters in Computer Science
Pennsylvania State University
Harrisburg, USA
nqr5356@psu.edu

Dr. Sayed Mohsin Reza
Assistant Professor of Computer Science
Pennsylvania State University
Harrisburg, USA
skr6024@psu.edu

Dr. Md Faisal Kabir
Assistant Professor of Computer Science
Pennsylvania State University
Harrisburg, USA
mpk5904@psu.edu

***Abstract*—Falls have been one of the leading causes for deaths in United States. Any insights about these falls can potentially inform interventions like education and awareness programs, exercise programs and home safety measures. The accumulation of clinical text data by healthcare organizations is growing, prompting a need for further exploration into the application of machine learning techniques on this data. This paper discussed about using state-of-the-art language model to extract concepts from the narratives. We propose a novel unsupervised approaches like Latent Dirchlet Allocation (LDA) and K-Means to classify the narratives into multiple classes of injury type. Additionally, we discussed human-assisted survey method to evaluate the performance of the models. The proposed model demonstrated the classification accuracy of 83% (LDA) and 75% (K-Means) on the unlabeled clinical narratives. The approach yielded valuable insights that might not be readily apparent through human observation. This study demonstrated the feasibility of utilizing unsupervised NLP methods to extract information from clinical narratives.**

***Index Terms*—Machine Learning, Natural Language Processing, Health Records, Unsupervised Learning, Medical Narratives, Clinical Narratives, Topic modelling, Finding Insights**

I. Introduction

Falls are one of leading cause of the injury-related deaths in older adults ($\geq$ 65 years). Nearly 165,005 older persons died as a result of falls in 2016-2020 time period; 261 million were treated in emergency rooms for fall-related injuries, with nearly 4.4 million of these patients being hospitalized. In 2014, 28.7% of elderly people reported falling, resulting in an estimated 29.0 million injuries [1]. Any insights about these falls can potentially inform interventions like Education and Awareness programs, Exercise programs and Home safety measures.

Electronic health record (EHR) is a digital version of patient's paper chart. EHRs are real-time, patient-centered records that make information available instantly and contain a patient's medical history, diagnoses, medications, treatment plans, immunization dates, allergies, radiology images, and laboratory and test results [2]. Consumer Product Safety Commission's (CPSC) National Electronic Injury Surveillance System (NEISS) program publicly hosts such health records of such fall-related injuries among older adults. NEISS collects the data from multiple hospitals' emergency departments. NEISS hospitals include significant inner-city trauma centers as well as large urban, suburban, rural, and children's hospitals. It uses a subset of these hospitals to get the data about the injuries [3]. We explore the potential of employing machine learning techniques to analyze these medical narratives, aiming to uncover insights that may elude human observation.

For many years, artificial intelligence (AI) and machine learning (ML) techniques have been used in healthcare. These advanced AI/ML tools have now become an important aspect of patient care in many cases for the health care providers. These tools are capable to assist with case triage and diagnosis [4], enhance image scanning and segmentation [5], support decision making, predict the risk of disease [6], [7] and in neuro-imaging [8], to determine the optimal sample size to test and utilize electronic records to eliminate database errors [9]. While many studies focus on AI/ML applications in healthcare, they predominantly center around medical imaging for predictions and diagnoses, leaving the textual components, such as medical narratives and other EHR aspects, under-explored.

The medical narratives are rich source of potential insights about how, when, and why people fall. Few studies [10], [11] employ rule-based supervision methods to categorize clinical narratives, but they typically limit the classification to only two categories. Given the complexity of the fall dataset, such a binary classification approach is insufficient. Moreover, these studies do not investigate alternative statistical techniques for generating weak labels in the training data. Antunes et al. [12] and Moen et al. [13] focus on exploring NLP techniques to determine the similarity between two medical words or sentences.

Nevertheless, there is a noticeable gap in the literature regarding the utilization of these similarity measures for further data extraction from clinical narratives. We propose a novel topic modeling approach in order to classify the narratives into multiple classes of injury type. On the other hand, we also process these topics more to uncover hidden knowledge about the falls in United States. This knowledge may offer a mechanism for leveraging unstructured clinical text data for characterization and monitoring of primary care practices and systems.

In this paper, we discuss the current research on medical narratives in II. In the following section III, a detailed explanation of the publicly available dataset and proposed natural language processing and machine learning techniques to create new insights on the clinical narratives is given. In the end, we show the results of the proposed methodology in section.

II. Literature Review

The Health Information Technology for Economic and Clinical Health Act of United States enforced in 2009 has fostered immense growth in use of Electronic Health Record (EHR). The number of healthcare institutions with a EHR systems has increased from 17% to 22% in 2009-10 [14]. One of the most challenges faced by researchers when using EHR data is to use the large amount of detailed patient information embedded in clinical narratives. Clinical narratives are often unlabeled due to their large size and resource-intensive labeling requirements. In the past decade, many studies have been published to use natural language processing (NLP) techniques to extract and model information embedded in such unlabeled clinical texts. While text classification is one of the sought-after application of clinical text mining. Many studies focus on other applications like Named Entity Recognition (NER), also known as Concept Extraction, weak/distant supervision, word/sentence embeddings, Topic Modelling and clustering. Next, we briefly discuss some of the mentioned techniques that are used in understanding and implementing the project.

An embedding is a mathematically meaningful vector representation of a word, sentence, or text, designed so that words or sentences with similar meanings or contexts have similar vectors. Similarity between two different embedding vectors is calculated using cosine similarity. In recent years, several word embedding models and pre-trained models [15], [16] which have been successfully applied in bio-medical field. However, these vectors have two fundamental limitations that render them unsuitable for bio-medical field. First, they use distinct vector to represent each word and do not consider the internal structure of the words. Second, these models struggle at learning rare or out of vocabulary words present in the training data.

Zhang et al. [17] create a new word embedding model which outperforms current state-of-the-art embedding models in all benchmarking tests. The authors train the model on PubMed Central (PMC) corpora and use medical subject headings (MeSH) and unified medical language system (UMLS) as biomedical domain knowledge. They construct MeSH term graph based on its RDF data and then sample the MeSH term embeddings. The fastText subword embedding model is used to learn the distributed word embeddings based on text sequences and MeSH term sequences. From clinical data mining standpoint, such embeddings not only excel in representing the irregularities of the medical text but also can incorporate structered information beyond what is found in text, proving beneficial for task like NER.

NER is the task of identifying and locating mentions of conceptual categories, in clinical texts, these are drug, symptom, or disease names. It is one of the most common extraction task that is run on the clinical texts. Simple NER systems are dictionary based which simply compare strings to a list of domain specific terms. Tulkens et al. [18] improve this strategy by proposing a composition fucntion for word representations of text and UMLS dictionary terms and subsequently evaluating using consine similarity. Stanza [19] offers state-of-the-art NER pipeline which are trained on biomedical [20], [21] and clinical domain database [22]. The authors train a forward and a backward LSTM character-level language model (CharLM) to get the word representation in each sentence which feeds a BiLSTM [23] POS tagger with conditional random decoder. We also explore spaCy's [24] NER model which is trained on BC5CDR dataset [21] to accomplish our NER tasks. One of the major advantage of NER systems is their flexibility in learning to tag different types of entities by swapping training datasets. However, clinical NER has the issue of domain specificity. The clinical narratives are complex and contain specialized medical terms, numerical scores and abbreviations which makes NER suffer from reduced accuracy.

Many of the machine learning approaches to clinical NLP generally suffer from the lack of structured training data. Most of the clinical texts are unstructured and unlabelled. This present two possible solution, weak and distant supervision. "Training machine learning models with weak or distant supervision approximates training with the true labels in terms of performance" [10]. Distant supervision involves using external knowledge base to label the data. An alternative is to use simple heuristic approaches like rule-based or statistical models (like clustering). Wang et al. [10] and Yao et al. [11] use keyword based weak labels to classify the texts and then deep learning models to classify the clinical narratives. Clustering algorithms like Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [25] has gained attention in recent years due to its ability to uncover complex and irregularly shaped clusters within data. Topic modeling techniques like Latent Dirchlet Allocation (LDA) [26], Latenet Semantic Indexing (LSI) [27] and varients of Fuzzy Latent Semantic Analysis (FLSA) [28] are recently gaining attention due to its adaptability to various types of texts and ability to discover hidden patterns in the documents by representing the documents in terms of topics.

While existing research emphasizes the application of unsupervised NLP techniques, particularly focusing on word embeddings and named entity recognition in clinical narratives, there remains a notable gap in comprehending the distinctive advantages that alternative unsupervised techniques, like topic modeling, might offer. Addressing this gap is pivotal for advancing our understanding of the latent structures within clinical narratives, thereby contributing to more nuanced and effective classification methodologies. This paper seeks to fill this gap by exploring the potential of topic modeling in classifying clinical narratives, thereby providing a comprehensive and innovative perspective in the landscape of healthcare NLP.

III. Methodology

In this section, we briefly discuss about the dataset that is used for our study. Further in the section, we describe the pre-process done to clean up the dataset. Finally, we explore different unsupervised NLP techniques used to extract hidden information from the clinical narratives and results.

A. *Dataset*

The data used in this study is a large, public dataset: the National Electronic Injury Surveillance System (NEISS). NEISS data is managed and maintained by the Consumer Product Safety Commission and come from a representative sample of emergency departments in the United States. It contains of approximately 115,000 confirmed cases of unintentional falls among the older adults. It contains various attributes describing the patient's demography, incident descriptions, treatment dates, clinical diagnoses, body parts affected etc which can be used as features for our study.

B. *Preprocess*

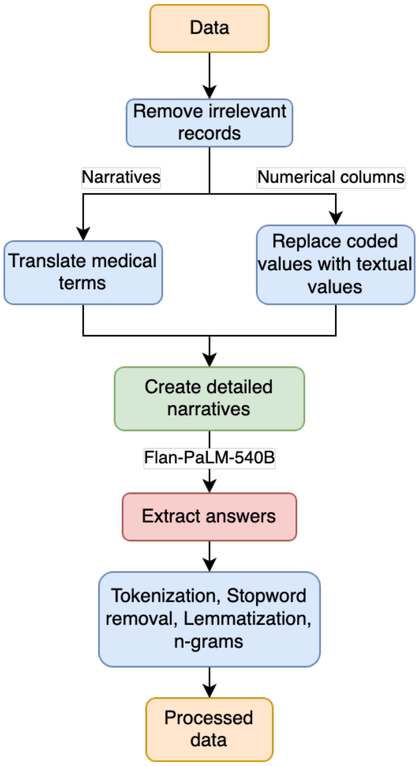

Data
Remove irrelevant records
Narratives
Numerical columns
Translate medical terms
Replace coded values with textual values
Create detailed narratives
Flan-PaLM-540B
Extract answers
Tokenization, Stopword removal, Lemmatization, n-grams
Processed data

Figure 1: Overview of the pre-processing applied on the medical narratives

The narratives are filled with many medical abbreviations which are widely used by the medical professionals in order to improve the efficiency without possibilities of obfuscation and ambiguity. A curated list of medical abbreviations relevant to the dataset is created from Wikipedia [29] as shown in table I. The abbreviations are then translated into their full form in order to improve the context in the narratives. Similarly, we also translate numerically encoded columns into textual columns. After the medical translation, we adopt the method proposed by Hegselmann et al. [30] to transform tabular data into fully textual data. To achieve this, we created a standard template of text such that all other important columns are accommodated along with the clinical narrative to create a lengthy and descriptive narrative as shown in Figure 2. We will refer to this lengthy and descriptive narrative as "detailed narrative" from hereafter in the paper. By using the information from other textual columns and feature values, detailed narratives often enable effective *zero-shot* or *few-shot* learning LLM models.

Abbreviation	**Full form**
dx	*diagnosis*
fx	*fracture*
pt	*patient*
biba	*brought in by ambulance*
bal	*blood alcohol level*
ams	*altered mental status*
tr	*trauma*

Table I: List of few medical abbreviations and their full forms that will be used to replace terms in narratives.

As proposed by Hegselmann et al. [30], we provide these detailed narratives to a pre-trained text-to-text transformer model like Google's T5 transformer. However, the challenge is to ask questions that are general enough for the model to return answers that align with the information we're trying to extract from the narratives like action, cause, diagnosis, how, what and where. We have specifically narrowed the questions to the following small set.

- What was the patient doing at the time of the incident? (*action*)
- What caused the patient to fall? (*cause*)
- What is the full diagnosis? (*diagnosis*)
- How did the fall occur? (*how*)
- What happened to the patient during the incident (*what*)
- Where did the fall occur? (*where*)

The T5 transformer model provides answers to these questions, each with a length of up to 20 words. The answers for each questions are saved under separate columns. One of caveat of this approach, is that some of these questions return similar or even same answers for certain narratives. However, it is better to have extra data than too few data. Another approach to avoid this issue is to train a model to generate questions that return specific answers but this would get computationally really expensive. So, we end up with above set of questions after improving them over multiple iterations after analyzing the answers after each iteration. The novelty in this approach lies in the transformer model's ability to extract information that might have been challenging due to the unstructured nature of clinical narratives.

The extracted answers are passed through a pipeline to enhance the quality of text representation for further tasks. The pipeline divides the sentences into individual words or subwords and eliminates any commonly occurring words like *the, is, it* etc. that do not contribute to the overall meaning of the text. These tokens are used to generate bigrams and trigrams capture contextual information and relationships between words. Finally, these words are lemmatized to ensure

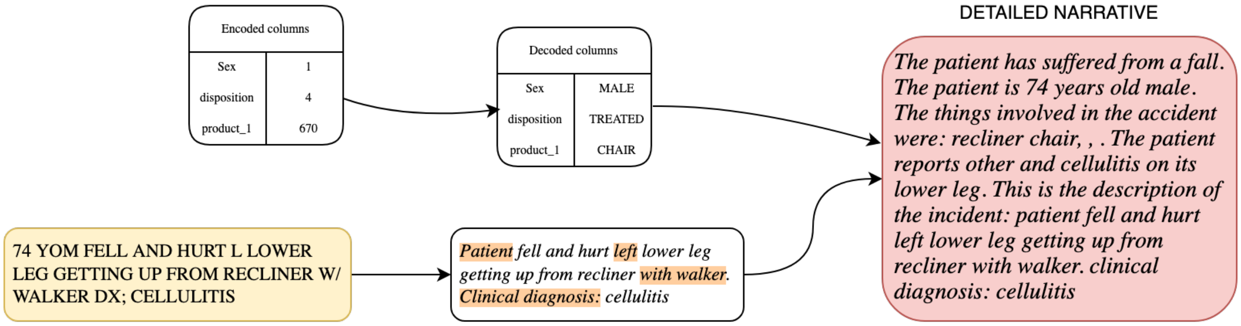

Encoded columns
Sex
1
disposition
4
product_1
670
Decoded columns
Sex
MALE
disposition
TREATED
product_1
CHAIR
DETAILED NARRATIVE
The patient has suffered from a fall. The patient is 74 years old male. The things involved in the accident were: recliner chair, , . The patient reports other and cellulitis on its lower leg. This is the description of the incident: patient fell and hurt left lower leg getting up from recliner with walker. clinical diagnosis: cellulitis
74 YOM FELL AND HURT L LOWER LEG GETTING UP FROM RECLINER W/ WALKER DX; CELLULITIS
Patient fell and hurt left lower leg getting up from recliner with walker. Clinical diagnosis: cellulitis

Figure 2: Overview of obtaining detailed narratives. We first translate the narratives to remove any medical abbreviations and then we decode the numerical columns. We combine translated narratives and decoded columns to create detailed narratives as shown in this figure.

the consistency in the analysis. The output of this pipeline is a list of tokens for each answer which will be used for topic modeling and classification of clinical narratives.

C. Topic Modeling

We use Latent Dirchlet Allocation (LDA) model which is an unsupervised probabilistic generative model of topics. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words [26]. According to Blei et al. [26], A k-dimensional Dirichlet random variable θ takes the dirchlet distribution. Given the parameters α and β, the joint distribution of a topic mixture θ, a set of N topics z, and a set of N words w is given by:

$$p(\theta,\, z, w \,|\, \alpha,\, \beta) = p(\theta,\, \alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta)$$

where $p(z_n|\theta)$ is simply θ_i for the unique i such that $z_n^i = 1$. Integrating over θ and summing over z, we obtain the marginal distribution of a document:

$$p(w \,|\, \alpha,\, \beta) = \int p(\theta,\, \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) \right) d\theta$$

Finally, taking the product of the marginal probabilities of single documents, we obtain the probability of a corpus:

$$p(D \,|\, \alpha,\, \beta) = \prod_{d=1}^{M} \int p(\theta_d,\, \alpha) \Gamma_d d\theta_d$$

where $\Gamma_d = \prod_{n=1}^{N} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta)$

Since the Dirchlet Distribution [31] is a continuous conditional probability distribution, Gibbs sampling [32] is applied on the distribution for digital computation. Gensim [33] is a python framework which provides easy-to-use implementation of Latent Dirchlet Allocation for topic modelling. We simply use gensim to create and train the LDA model on the answers that were obtained from T5 transformer model. For each question column, we create a term-document matrix which represents the frequency of terms (words or phrases) in documents in a collection. This term-document matrix will be the input of the LDA model. Apart from the term-document matrix, the LDA model also takes a lot of hyperparameters like number of topics, trigram and bigram count, chunksizes and passes. We run optuna [34] framework on a ranges of values create an optimal training model. Table II shows optimal hyperparameters for each column.

Column	**Num Topics**	**Chunk size**	**passes**	**Bigram count**	**Trigram count**
action	10	1500	4	1	100
cause	10	1000	1	4	70
diagnosis	10	500	1	1	100
how	70	500	4	4	10
what	60	500	1	7	70
where	10	1500	1	4	10

Table II: Hyperparameters for Latent Dirchlet Allocation (LDA) obtained after going through 40 trails per column using optuna framework. The hyperparameters are optimized on 30000 rows. The LDA model is trained with these optimized hyperparameters on full dataset.

D. Clustering

We also evaluate the potential of several clustering models as a baseline to assess the performance of LDA. Specifically, we select K-Means which offers us the flexibility of chosing number of clusters needed. Whereas, DBSCAN and HDBSCAN variants change the number of clusters based on the data distribution and other hyperparameters. For clustering, we use word2vec [35] to convert the answer tokens to vector representations. The word2vec model takes many parameters like vector size, number of windows, skip-grams etc. We optimize the hyperparameters by running a parameter grid search evaluated on the similarity of two very close words. The word2vec model is trained on the entire corpus of answer tokens including the keyword tokens which will be discussed in section III-E. Now, we combine the tokens from ‘diagnosis‘ question with keyword tokens and provide all these tokens to the clustering model to form 8 clusters, ideally, one for each class. The hyperparameters of the models are also optimized using optuna [34] based on Silhouette Score, Calinski-Harabasz Index and Davies-Buldin Index.

E. Classification

We classify the narratives into 8 different injury types which cover most common injury types. Each class is assigned with a keyword as specified by Hedegaard et al. [36] in their Center for Disease Control (CDC) publication. Table III provide the keywords used for each injury class.

Injury class	Keyword(s)
0	fracture
1	sprain
2	strain
3	laceration
4	contusion, concussion
5	dislocation
6	hematoma, hemorrhage
7	injury, other

Table III: Class vs Keyword. Each class gets a particular keyword(s).

These keywords are then added to the training corpus of all the models as different case file. All the narratives which have been grouped in the same cluster as these injury keywords are labelled with injury class associated with the keyword. A visual representation of the classification pipeline is depicted in Figure 3.

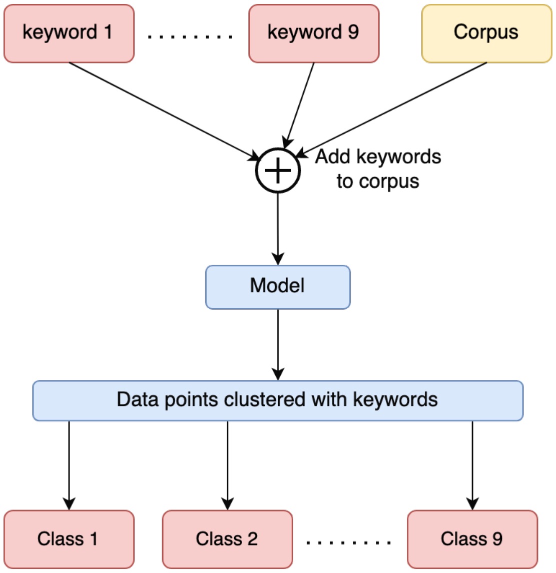


Figure 3: Overview of classification. Every injury class is associated with a distinct keyword, recommended by the CDC [36]. These keywords are incorporated into the model's training corpus. Data points clustered or grouped with these keywords are classified as same injury type as the keywords.

F. Evaluation

The evaluation of the models depends on the evaluation goals, which are operationalized through various metrics. The performance of LDA model is evaluated based on the perplexity and coherence metric. Perplexity captures how surprised a model is of new data it has not seen before, and is measured as the normalized log-likelihood of a held-out test set. On the other hand, coherence measures score a single topic by measuring the degree of semantic similarity between high scoring words in the topic. Similarly, for the clustering models we use Silhouette Score, Calinski-Harabasz Index and Davies-Buldin Index. Silhouette Score measures how well-separated clusters are. Calinski-Harabasz Index measures the ratio of between-cluster variance to within-cluster variance. Davies-Bouldin Index measures compactness and separation between clusters.

The ideal way for evaluating the classification results of unlabeled dataset is through human evaluation of models' performance. To conduct human evaluation, we compiled a test sample comprising 200 narratives, presenting five clinical narratives to each of 20 individuals and requesting them to select the appropriate injury class. We then compare the prediction of the models with the human evaluation results. This will provide us the information about the performance of the models. We subsequently compare the model predictions with the human evaluation results, offering insights into the models' performance.

IV. Results

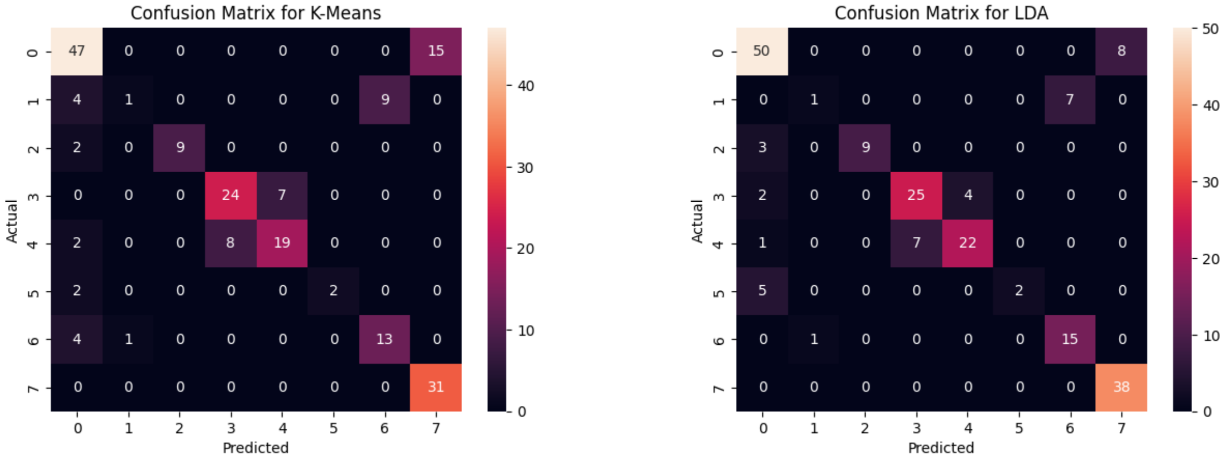
Table IV presents a comprehensive breakdown of the class-wise accuracy achieved by both models in categorizing injury classes within medical narratives. K-Means exhibits an overall accuracy of 72%, while LDA outperforms with an accuracy of 81%.

Class	K-Means	LDA
0	0.87	0.93
1	0.63	0.67
2	0.75	0.87
3	0.59	0.75
4	0.78	0.83
5	0.77	0.81
6	0.67	0.87
7	0.73	0.78
Overall	0.72	0.81

Table IV: Class-wise accuracy provided by K-Means and LDA models

Figure 5 presents a comprehensive comparison of the class-wise F1 scores obtained by both models. Notably, both models exhibit suboptimal performance in the classification of class 1, specifically related to sprains. A plausible explanation for the lower F1 score in this class could be attributed to the inadequacy of data size. The insufficient representation of instances belonging to class 1 in the dataset might impede the models' ability to capture and generalize patterns effectively, leading to reduced performance of the model. In general, LDA outperforms the K-Means clustering algorithm.

Figure 4 illustrates the confusion matrix for these models. Notably, an interesting pattern emerges wherein clinical narratives belonging to fracture cases are consistently misclassified as other types of injuries in both models. This misclassification suggests the presence of substantial irrelevant data in the dataset. Consequently, further data preprocessing is necessary to enhance the models' performance by eliminating



	0	1	2	3	4	5	6	7
0	47	0	0	0	0	0	0	15
1	4	1	0	0	0	0	9	0
2	2	0	9	0	0	0	0	0
3	0	0	0	24	7	0	0	0
4	2	0	0	8	19	0	0	0
5	2	0	0	0	0	2	0	0
6	4	1	0	0	0	0	13	0
7	0	0	0	0	0	0	0	31

(a) Confusion matrix for K-Means clusturing algorithm

	0	1	2	3	4	5	6	7
0	50	0	0	0	0	0	0	8
1	0	1	0	0	0	0	7	0
2	3	0	9	0	0	0	0	0
3	2	0	0	25	4	0	0	0
4	1	0	0	7	22	0	0	0
5	5	0	0	0	0	2	0	0
6	0	1	0	0	0	0	15	0
7	0	0	0	0	0	0	0	38

(b) Confusion matrix for Latent Dirchlet Allocation algorithm

Figure 4: Confusion matrices

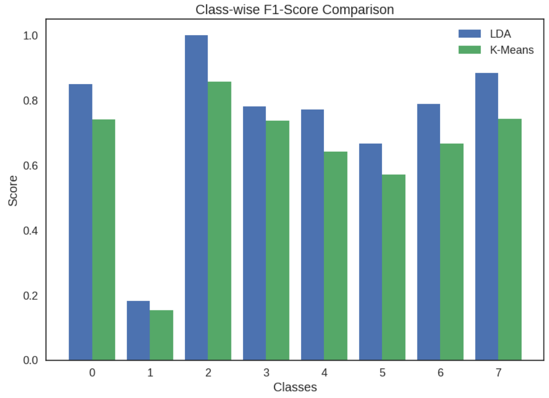


Figure 5: Plot of class-wise F1 scores.

this irrelevant data, ensuring more accurate and reliable injury classification.

V. Conclusion

In this study, we present a novel approach of classifying clinical narratives into different injury types and improving insights on the data. To improve the quality of the data we explore a pre-trained text-to-text transformer, FLan-PaLM-540B, to extract particular incident related information. We discuss in detail unique strengths and weaknesses owing to their underlying statistical properties of unsupervised learning techniques like Latent Dirchlet Allocation and K-Means paired with Word2Vec in order to tag or classify the clinical narratives and methods to evaluate the performance. As the adoption of EMRs increases and health care organizations amass increasingly large volumes of clinical text data, such insights may offer a mechanism for leveraging unstructured clinical text data for characterization and monitoring of primary care practices and systems. Additionally, these insights can potentially inform interventions like education and awareness programs, exercise and safety programs.

Despite notable challenges in handling specific injury types, future efforts should concentrate on refining or finding models to ensure generalized learning across all classes. Addressing limitations in certain injury types is crucial for the robustness of the classification system. Additionally, exploring automated evaluation mechanisms using advanced language models like GPT, BERT, or similar technologies holds promise for enhancing the assessment process.

References

[1] Falls and fall injuries among adults aged above 65 years - united states, 2014. [Online]. Available: https://www.cdc.gov/injury/features/older-adult-falls/index.html

[2] H. Zheng and S. Jiang, "Frequent and diverse use of electronic health records in the united states: A trend analysis of national surveys." *Digit Health*, vol. 8, p. 20552076221112840, Jan-Dec 2022.

[3] National electronic injury surveillance system - all injury program (neiss-aip). [Online]. Available: https://health.gov/healthypeople/objectives-and-data/data-sources-and-methods/data-sources/national-electronic-injury-surveillance-system-all-injury-program-neiss-aip

[4] Z. Liang, G. Zhang, J. X. Huang, and Q. V. Hu, "Deep learning for healthcare decision making with emrs," in *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2014, pp. 556–559.

[5] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017. [Online]. Available: https://doi.org/10.1038/nature21056

[6] B. P. Nguyen, H. N. Pham, H. Tran, N. Nghiem, Q. H. Nguyen, T. T. Do, C. T. Tran, and C. R. Simpson, "Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records," *Computer Methods and Programs in Biomedicine*, vol. 182, p. 105055, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016926071930327X

[7] Y. Ge, Q. Wang, L. Wang, H. Wu, C. Peng, J. Wang, Y. Xu, G. Xiong, Y. Zhang, and Y. Yi, "Predicting post-stroke pneumonia using deep neural network approaches," *International Journal of Medical Informatics*, vol. 132, p. 103986, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1386505618312723

[8] M. Faturrahman, I. Wasito, N. Hanifah, and R. Mufidah, "Structural mri classification for alzheimer's disease detection using deep belief network," in *2017 11th International Conference on Information and Communication Technology and System (ICTS)*, 2017, pp. 37–42.

[9] C. Rudin and B. Ustun, "Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice," *Interfaces*, vol. 48, 10 2018.

[10] Y. Wang, S. Sohn, S. Liu, F. Shen, L. Wang, E. J. Atkinson, S. Amin, and H. Liu, “A clinical text classification paradigm using weak supervision and deep representation.” *BMC Med Inform Decis Mak*, vol. 19, no. 1, p. 1, Jan 2019.

[11] L. Yao, C. Mao, and Y. Luo, “Clinical text classification with rule-based features and knowledge-guided convolutional neural networks.” *BMC Med Inform Decis Mak*, vol. 19, no. Suppl 3, p. 71, Apr 2019.

[12] R. Antunes, J. a. F. Silva, and S. Matos, “Evaluating semantic textual similarity in clinical sentences using deep learning and sentence embeddings,” in *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, ser. SAC ’20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 662–669. [Online]. Available: https://doi.org/10.1145/3341105.3373987

[13] H. Moen, L.-M. Peltonen, M. Koivumäki, H. Suhonen, T. Salakoski, F. Ginter, and S. Salanterä, “Improving layman readability of clinical narratives with unsupervised synonym replacement.” *Stud Health Technol Inform*, vol. 247, pp. 725–729, 2018.

[14] W. H. Henricks, “”meaningful use” of electronic health records and its relevance to laboratories and pathologists.” *J Pathol Inform*, vol. 2, p. 7, Feb 2011.

[15] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in Neural Information Processing Systems*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., vol. 26. Curran Associates, Inc., 2013. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

[16] J. Pennington, R. Socher, and C. Manning, “GloVe: Global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: https://aclanthology.org/D14-1162

[17] Y. Zhang, Q. Chen, Z. Yang, H. Lin, and Z. Lu, “Biowordvec, improving biomedical word embeddings with subword information and mesh,” *Scientific Data*, vol. 6, no. 1, p. 52, 2019. [Online]. Available: https://doi.org/10.1038/s41597-019-0055-0

[18] S. Tulkens, S. Šuster, and W. Daelemans, “Unsupervised concept extraction from clinical text through semantic composition,” *Journal of Biomedical Informatics*, vol. 91, p. 103120, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046419300383

[19] Y. Zhang, Y. Zhang, P. Qi, C. D. Manning, and C. P. Langlotz, “Biomedical and clinical English model packages for the Stanza Python NLP library,” *Journal of the American Medical Informatics Association*, 06 2021.

[20] S. Pyysalo and S. Ananiadou, “Anatomical entity mention recognition at literature scale.” *Bioinformatics*, vol. 30, no. 6, pp. 868–875, Mar 2014.

[21] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, and Z. Lu, “Biocreative v cdr task corpus: a resource for chemical disease relation extraction.” *Database (Oxford)*, vol. 2016, 2016.

[22] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, “2010 i2b2/va challenge on concepts, assertions, and relations in clinical text.” *J Am Med Inform Assoc*, vol. 18, no. 5, pp. 552–556, Sep-Oct 2011.

[23] Z. Huang, W. Xu, and K. Yu, “Bidirectional LSTM-CRF models for sequence tagging,” *CoRR*, vol. abs/1508.01991, 2015. [Online]. Available: http://arxiv.org/abs/1508.01991

[24] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd. (2020) spaCy: Industrial-strength Natural Language Processing in Python.

[25] C. Malzer and M. Baum, “Hdbscan: An alternative cluster extraction method for HDBSCAN,” *CoRR*, vol. abs/1911.02282, 2019. [Online]. Available: http://arxiv.org/abs/1911.02282

[26] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, no. null, pp. 993–1022, mar 2003.

[27] M. A. Ahmad, C. Eckert, and A. Teredesai, “Interpretable machine learning in healthcare,” in *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, 2018, pp. 559–560.

[28] E. Rijcken, F. Scheepers, P. Mosteiro, K. Zervanou, M. Spruit, and U. Kaymak, “A comparative study of fuzzy topic models and lda in terms of interpretability,” in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2021, pp. 1–8.

[29] Wikipedia contributors, “List of medical abbreviations — Wikipedia, the free encyclopedia,” https://en.wikipedia.org/w/index.php?title=List_of_medical_abbreviations&oldid=1177253313, 2023, [Online; accessed 10-November-2023].

[30] S. Hegselmann, A. Buendia, H. Lang, M. Agrawal, X. Jiang, and D. Sontag, “Tabllm: Few-shot classification of tabular data with large language models,” 2023.

[31] Wikipedia contributors, “Dirichlet distribution — Wikipedia, the free encyclopedia,” https://en.wikipedia.org/w/index.php?title=Dirichlet_distribution&oldid=1170494520, 2023, [Online; accessed 11-November-2023].

[32] ——, “Gibbs sampling — Wikipedia, the free encyclopedia,” https://en.wikipedia.org/w/index.php?title=Gibbs_sampling&oldid=1171763396, 2023, [Online; accessed 11-November-2023].

[33] R. Rehurek and P. Sojka, “Gensim–python framework for vector space modelling,” *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, vol. 3, no. 2, 2011.

[34] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A next-generation hyperparameter optimization framework,” 2019.

[35] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” 2013.

[36] H. Hedegaard, R. L. Johnson, M. F. Garnett, and K. E. Thomas, “The 2020 international classification of diseases, 10th revision, clinical modification injury diagnosis framework for categorizing injuries by body region and nature of injury.” *Natl Health Stat Report*, no. 150, pp. 1–27, Dec 2020.