MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231040

Rule-Based Text Classification of Dental Diagnosis

Mei WANG[a,1] Anushka AGRAWAL[b] Nicole ROGERS[a] Vanchit JOHN[a] Thankam THYVALIKAKATH[a,c]
[a] *Indiana University School of Dentistry*
[b] *Indiana University School of Informatics and Computing*
[c] *Center for Biomedical Informatics, Regenstrief Institute, Inc.*

ORCiD ID: Mei Wang https://orcid.org/ 0000-0002-6585-191X

Abstract. Unstructured medical records boast an abundance of information that could greatly facilitate medical decision-making and improve patient care. With the development of Natural Language Processing (NLP) methodology, the free-text medical data starts to attract more and more research attention. Most existing studies try to leverage the power of such unstructured data using Machine Learning algorithms, which would usually require a relatively large training set, and high computational capacity. However, when faced with a smaller-scale project, opting for an alternative approach may be more effective and practical. This project proposes an efficient and light-weight rule-based approach to categorize dental diagnosis data. It not only fills the void of dental records in the medical free-text processing area, but also demonstrates that with expertly designed research structure and proper implementation, simple method could achieve our study goal very competently.

Keywords. Dental Texts, Rule-Based Classification, Natural Language Processing

1. Introduction

Unstructured medical data has been gaining increasing attention in healthcare settings. With the evolution of various Natural Language Processing (NLP) models, information underneath the free text is revealed and interpreted to facilitate medical decision-making and improve patient care. Current NLP methods largely consist of machine learning algorithms, which typically have prerequisites for the dataset size and computing capacity. With the ascendance of various machine learning NLP models, we should never overlook the efficiency and practicality of other text-processing approaches.

This study showcases a non-AI text classification project. The goal is to categorize dental diagnosis texts into three major categories to understand the periodontal health status of dental patients.

[1] Corresponding Author: Mei Wang, wangmei@iu.edu; Phone: +1(317)-274-7130; 415 Lansing Street Indianapolis, IN 46202-2876

2. Methods

This text classification project is a component of a larger project that is studying the clinical training experiences for graduation that predoctoral students receive at Indiana University School of Dentistry (IUSD).

IUSD currently utilizes a requirement-based teaching model: in periodontics, for example, requirements on the numbers and types of procedures must be fulfilled before a predoctoral student could graduate. Therefore, categorizing the periodontal diagnosis texts helps define the periodontal disease profile of the IUSD predoctoral patients, ultimately contributing to an understanding of the learning accomplished by IUSD predoctoral students.

2.1. About the dataset

Unlike other medical records, dental clinical texts are rarely studied or utilized in existing research. Currently, the work of Patel et al. [3] is the only work that includes analysis of dental texts, with the attempt to automate the diagnosis of gingivitis and periodontitis.

The dataset of this project includes all patients with at least one completed procedure from July 1, 2016 through June 30, 2020 as the study cohort. Among this cohort, patients with at least one periodontal treatment will have a periodontal treatment plan form saved in IUSD's electronic dental record (axiUm) database. On this form, there is a free-text field named "Diagnosis", where the provider typically inputs one sentence summarizing the patient's periodontal health situation.

14,532 patients' periodontal diagnosis texts were extracted as the dataset to be classified. Some examples of the diagnosis texts are shown below:

- "Generalized mild chronic periodontitis"
- "Healthy gingiva on an intact periodontium"
- "Dental plaque associated gingivitis on a reduced periodontium on a successfully treated periodontal patient with Peri implant Health #19"

The complete dataset had 88,869 rows. Each row was a complete diagnosis text answer. The dataset description is as follows:

- Total word count: 208,096
- Total unique sentence/row count: 24,065
- Max. word/sentence: 129
- Min. word/sentence: 1
- Average word/sentence: 8.74

2.2. Classification structure

After carefully reviewing the texts and consulting with the providers, we decided to classify them into three major categories: gingival health, gingivitis, and periodontitis. Periodontitis is further sub-categorized into Stage 1, Stage 2 Stage 3, and Stage 4. We relied on the providers' expertise to devise the classification rules. For each category/sub-category, we reviewed sufficient examples of diagnosis texts. Then we identified strings

unique to that particular group of texts, which were the keywords that defined that category. The classification was finalized through keyword matching, and labels were assigned accordingly.

2.3. Preprocessing and classification

The dataset was free-text data. Therefore, multi-step pre-processing was needed before it could be used for text classification. The typical procedures of pre-processing were performed following the workflow below (Figure 1).

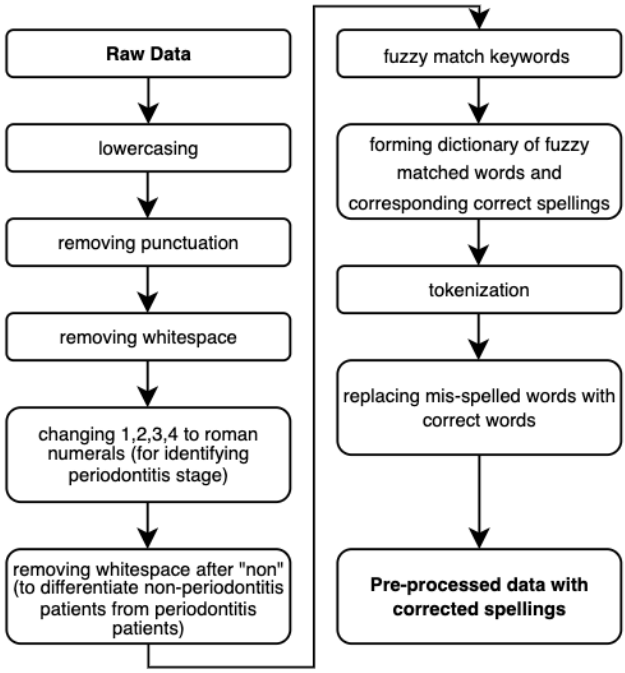

Figure 1. Raw data pre-processing workflow.

To increase the success rate of matching, fixing keyword typographical errors is one of the most important issues to be addressed during pre-processing. Before pre-processing, an initial investigation revealed more than 500 unique sentences that contain misspelled keywords, which could not be categorized. To fix this problem, the Python package "Fuzzy Wuzzy" was used to realize any approximate matches of keywords.

Thanks to the meticulously designed classification structure, we could extract unique keyword strings for each category/subcategory. Therefore, the rule-based classification could be accomplished by keyword match. The keyword strings used to assign category labels with some example diagnosis texts are shown in Figure 2 below.

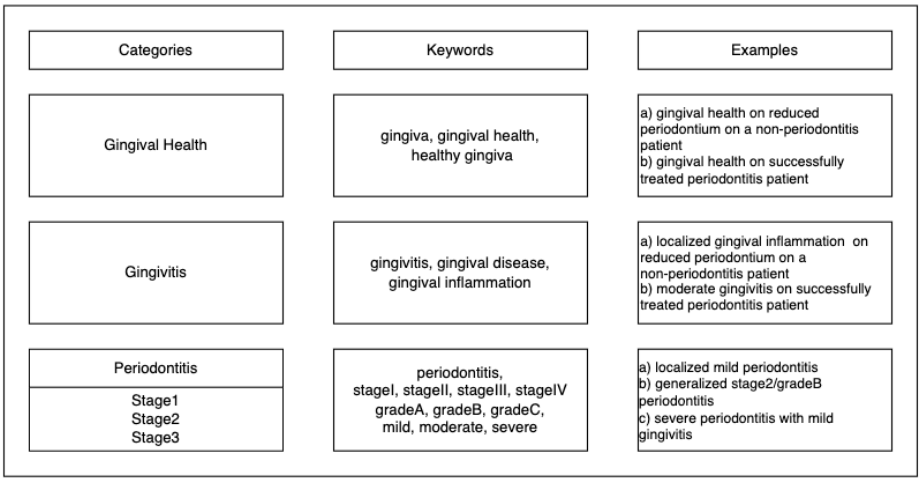

Figure 2. Classification structure, keywords, and examples.

The severity of the periodontal conditions increased from top to bottom as shown in Figure 2. When one record belongs to multiple categories, the more severe diagnosis category was taken and assigned.

3. Results

This method effectively classified most of the diagnosis texts. As a result, only 123 sentences were uncategorized, and 130 periodontitis records with no stage information, which meant this method classified 99% of the data. The distribution of the categories is shown below (Figure 3).

4. Discussion

This project reminds us that for NLP models, machine learning methods are not always better. Sometimes, small is beautiful. Based on a profound understanding of the data, the text data can be classified by a carefully designed set of rules. The results demonstrate that the proper implementation of a simpler approach, combined with an expertly devised analysis structure, is very powerful and effective.

5. Conclusions

To conclude, for medical NLP tasks, although machine learning algorithms are attracting most research attention, we should not overlook simpler approaches. Backed by in-depth expertise, and supported by carefully designed analysis structure, simple rule-based approach could perform excellently. It is also worth noting that the remarkable performance of this approach could not be realized without the high-standard training the IUSD providers receive, where they are required to use a fixed vocabulary to fill the

diagnosis field. That ensures the success of defining each category/sub-category using the keyword strings.

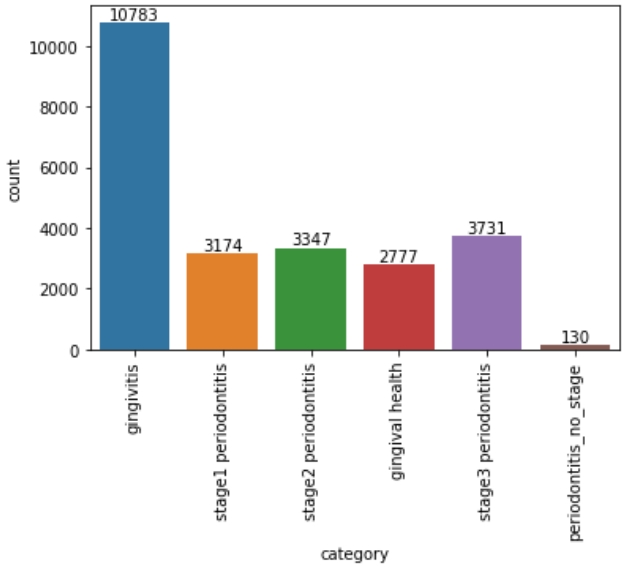

Figure 3. Histogram of classification results.

References

[1] Al-Doulat A, Obaidat I, Lee M. Unstructured medical text classification using linguistic analysis: a supervised deep learning approach. In: Proceedings of 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA); 2019 Nov 03-07; Abu Dhabi, United Arab Emirates: IEEE; c2020. doi:10.1109/aiccsa47632.2019.9035282.

[2] Bae YS, Kim KH, Kim HK, Choi SW, Ko T, Seo HH, Lee HY, Jeon H. Keyword extraction algorithm for classifying smoking status from unstructured bilingual electronic health records based on natural language processing. Applied Sciences. 2021 Sep;11(19):8812, doi:10.3390/app11198812.

[3] Patel JS, Kumar K, Zai A, Shin D, Willis L, Thyvalikakath TP. Developing automated computer algorithms to track periodontal disease change from longitudinal electronic dental records. Diagnostics. 2023 Mar;13(6):1028, doi:10.3390/diagnostics13061028.

[4] Villena J, Collada-Pérez S, Serrano SL, Gonzalez-Cristobal J. Hybrid approach combining machine learning and a rule-based expert system for text categorization. In: Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference; 2011 May 18-20; Palm Beach, FL.

[5] Yao L, Mao C, Luo Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. In: Vydiswaran VGV, Zhang Y, Wang Y, Xu H, editors. Proceedings of the first International Workshop on Health Natural Language Processing (HealthNLP 2018); 2018 Jun 4-7; New York, NY. New York (NY); c2019. p.71, doi:10.1186/s12911-019-0781-4.

[6] Yousef M, Voskergian D. Textnettopics: text classification based word grouping as topics and topics' acoring. Frontiers in Genetics. 2022 Jun;13:893378, doi:10.3389/fgene.2022.893378.