**Enhancing Clinical Analysis through NLP Integration**

JANAKI RAMYA NAMBURU

Dept. of Health Informatics

School of Health Professions

Rutgers University

Newark, USA

[jn746@shp.rutgers.edu](mailto:jn746@shp.rutgers.edu)

**Abstract:**

**Background:** NLP has become increasingly prevalent in healthcare analytics, offering a powerful approach to extract valuable insights from unstructured medical text data. This study aims to enhance clinical analysis by investigating smoking status classification and the association between alkaline phosphatase (ALP) and bone within discharge summaries. **Methods:** Using a dataset of 398 discharge summaries, we employed Naïve Bayes classifiers trained on different n-gram models to ascertain smoking status. Additionally, a Spacy-based system was utilized for Named Entity Recognition (NER) and relation extraction between ALP and bone. **Results:** The bigram-based Naïve Bayes classifier achieved the highest accuracy at 65%. Despite successful entity extraction, accurate associations between ALP and bone were difficult to establish, with all extracted relations being false. The system achieved a 70% correct extraction rate for ALP values. These findings highlight both the potential and limitations of NLP techniques in clinical analysis. **Conclusion:** NLP techniques exhibit promise in healthcare, facilitating tasks such as smoking status classification and entity extraction. However, challenges persist in relation extraction and value extraction, necessitating further refinement for enhanced clinical analysis.

**Keywords:**

Natural Language Processing, Clinical Analysis, Smoking Status Classification, Named Entity Recognition, Relation Extraction, Alkaline Phosphatase, Bone

**Introduction:**

Natural language processing (NLP) is “Any computer-based algorithm that handles, augments, and transforms natural language so that it can be represented for computation.” Nowadays, NLP is the widely used analytical technique in the healthcare industry, particularly for "big data" analysis. Although the medical language is often ambiguous and complex, NLP has been successfully applied in healthcare for different purposes, such as recognizing risk factors of medical conditions, evaluating the efficiency of care and costs, and extracting insights from the unstructured text of medical records.[1]

The healthcare industry faces a huge challenge with the burden of chronic diseases such as cancer, diabetes, and hypertension. Despite the significant progress made in finding new treatments and prevention strategies, the incidence of these diseases is still increasing. One promising novel approach would be the utilization of medical health records to analyze patient data, conduct medical research, and improve clinical decision-making.[2]

NLP techniques have proven to be valuable tools in managing the vast amount of information present in the health and medical field. These techniques aid in various tasks such as aggregating and summarizing patient notes, analyzing treatments, extracting and retrieving information from extensive discharge summaries, and understanding the semantics of patient queries. Furthermore, NLP can assist in medical decision-making by automatically analyzing similarities and distinctions among large volumes of textual data, thereby providing recommendations for appropriate actions, thus aiding domain experts in their decision-making processes.[3]

Smoking is the leading preventable cause of death. It is the responsibility of healthcare professionals to help patients quit smoking, as it is a crucial aspect of treating many diseases such as lung cancer and coronary heart disease.[4] In the United States, almost 500,000 deaths each year are caused by smoking. Effective preventive strategies depend on the identification of tobacco-dependent individuals.[5] Alkaline phosphatase is a marker of bone formation and bone turnover and is used in the evaluation of skeletal status.[6] The objective of this study is to utilize a Naïve Bayes (NB) classifier trained on different n-gram models to ascertain smoking status in discharge summaries, followed by an evaluation of the system's performance and also to develop and evaluate a Spacy-based system for automatically extracting instances of alkaline phosphatase (ALP) and bone from medical text and analyzing the association between them using information retrieval and relation extraction techniques.

**Methodology:**

The dataset was obtained from the “n2c2 NLP Research Data Sets of Unstructured notes from the Research Patient Data Registry at Partners Healthcare” from the DBMI data portal of Harvard Medical School.[7] This dataset was created at a former “NIH-funded National Centre for Biomedical Computing (NCBC) known as i2b2: Informatics for Integrating Biology and the Bedside.” The dataset consists of 399 observations/ discharge summaries with patient ID, smoking status, and free text of discharge summary. Patients are classified based on smoking status into 5 categories: smoker, current smoker, past smoker, nonsmoker, and unknown. Text preprocessing steps like removal of null values, duplicates, punctuations, special characters, stop words, lower casing of text, and lemmatization were done to improve the outcome. Performing stemming on the processed texts was avoided since it might change the meaning of words. The dataset was split for the training and testing phase (0.80, 0.20). A Naïve Bayes classifier was used which is trained on features extracted from unigrams, bigrams, and trigrams to perform text classification.

For Named Entity Recognition (NER) and relation extraction between alkaline phosphatase (ALP) and bone mentions, a Spacy-based system was devised. The system utilized the PhraseMatcher for information retrieval, identifying instances of ALP and bone mentions in the medical text data. Additionally, the DependencyMatcher was employed for relation extraction, capturing the syntactic dependencies between ALP and bone entities. Furthermore, a rule-based function was applied to each text to verify the existence of a relation between ALP and bone, serving as a complementary method to the automated extraction process. Additionally, another rule-based function was implemented to extract the values of ALP mentioned in the context, providing quantitative information for further analysis. Lastly, context windows surrounding occurrences of ALP and bone were extracted to furnish additional context for comprehensive analysis. This combined approach aimed to enhance the accuracy and comprehensiveness of the association between ALP and bone extracted from the medical text data, facilitating deeper insights into the relationship between these entities in the health domain.

**Results:**

**Text classification:**

Unigram-based Classifier:

This classifier yielded an overall accuracy of 20% (Table 1). Despite its relatively low accuracy, it demonstrated notable recall for the "Past Smoker" category (83%). However, precision and recall for other categories, such as "Current Smoker" and "Unknown," were suboptimal, suggesting challenges in accurately identifying these groups based on individual words. However, the low recall indicates that many instances of "Unknown" smoking status were missed by the classifier.

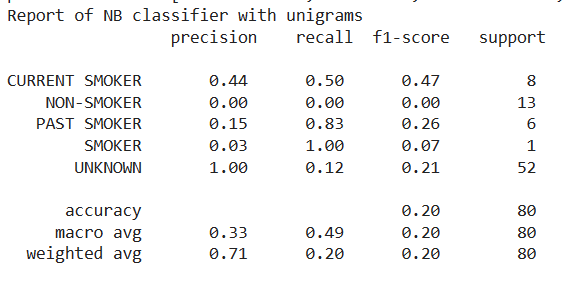
****

Table 1 - NB classifier – Unigrams

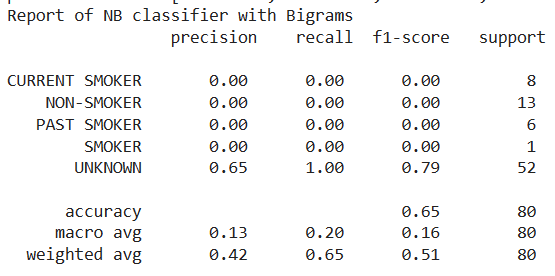
****

Table 2 - NB classifier- Bigrams

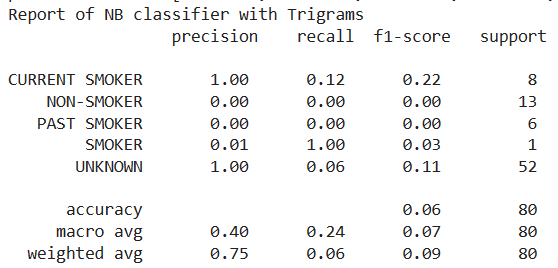
****

Table 3 - NB classifier – Trigrams

Bigram-based Classifier

The bigram-based classifier achieved an accuracy of 65% (Table 2), representing a notable improvement compared to the unigram-based approach. The precision and recall for the "Unknown" category notably improved, indicating better performance in identifying this group. The bigram-based classifier failed to correctly classify any instances for the categories Current Smoker, Non-Smoker, Past Smoker, and Smoker resulting in zero precision and recall for each.

Trigram-based Classifier

The trigram-based classifier demonstrated the lowest accuracy among the three approaches, with an accuracy of 6% (Table 3). While the precision for the "Unknown" category was perfect, the overall performance was severely limited.

**Named- Entity Recognition and Relation Extraction:**

The NER process successfully extracted entities of alkaline phosphatase and bone from the medical text data without encountering significant challenges. However, despite efforts, the system faced difficulties in identifying relations between "alkaline phosphatase" and "bone" mentions, with all extracted relations being false. This might be due to the complexities in syntactic structures, contextual nuances, and variations in language usage across medical records. While attempting to extract ALP values, the system achieved a correct extraction rate of 70% (7 out of 10 instances). The challenges in value extraction primarily stemmed from inconsistencies in formatting, variations in numerical representation, and the presence of noise or irrelevant text surrounding ALP mentions.

**Visualizations:**

Figure 1 illustrates a notable disparity in the distribution of smoking labels, with the "unknown" category substantially outnumbering other labels. This skewed distribution suggests a prevalence of missing or unclassified smoking statuses within the dataset, likely influencing the text classification model's performance and contributing to the observed lower precision scores.  
In Figure 2, we can observe the word cloud with word frequencies. Each word's size in the cloud correlates with its occurrence frequency in the text, with larger words denoting higher frequencies. Notably, we observe pivotal terms recurrent in medical records, such as “patient”, “history”, “mg”, “time”, “admitted”, “Discharge” and “Dr”. This visualization makes it easy to spot the most important words in the dataset, facilitating a quick grasp of key themes and topics in the dataset.

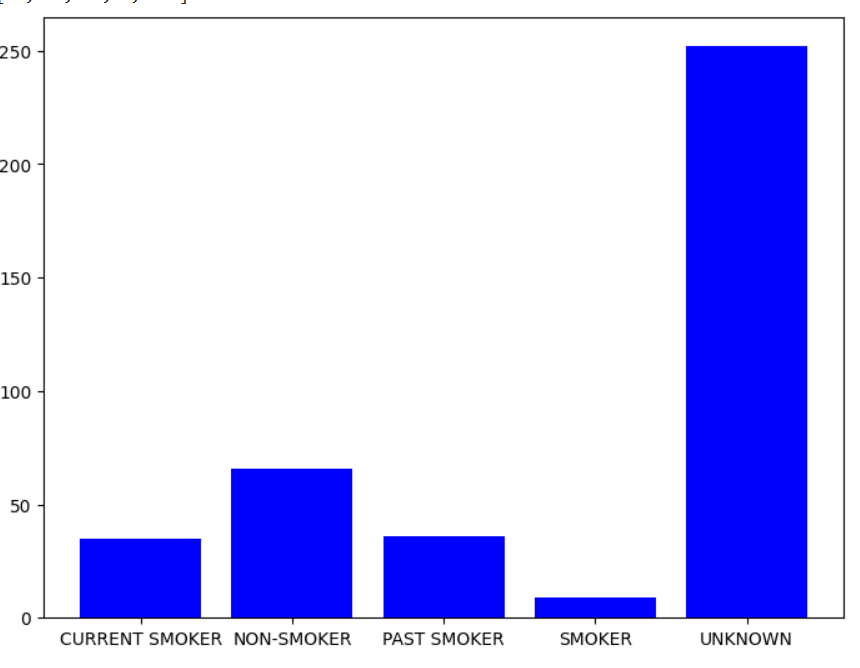


Figure 1 Distribution of Smoking Labels

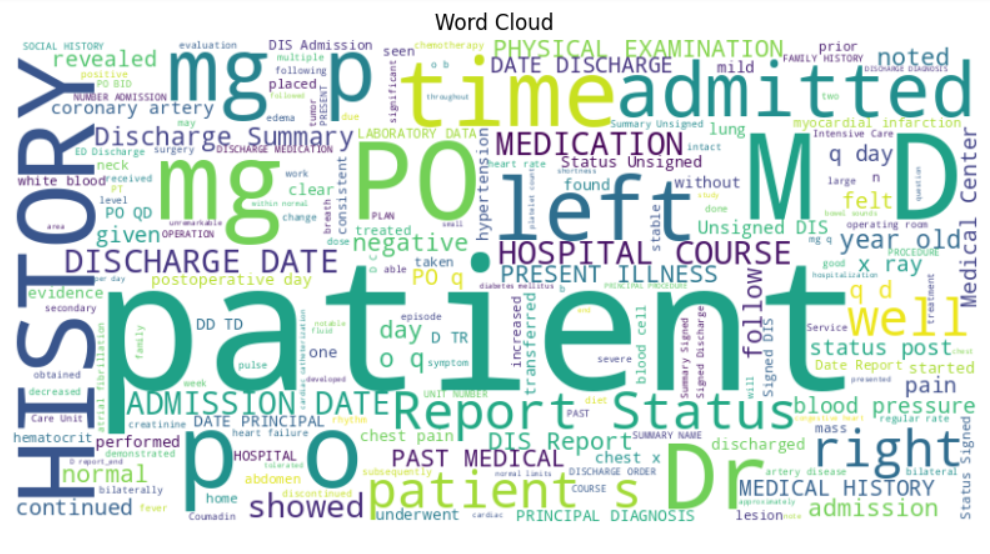
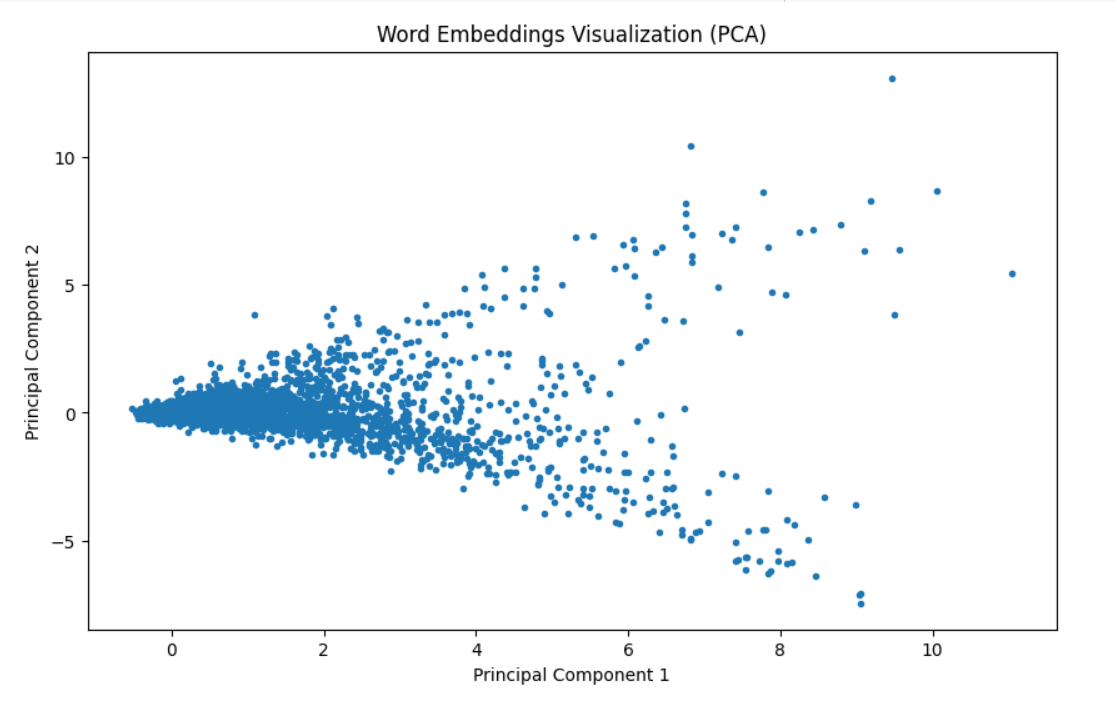


Figure 2 – Word Cloud



*Figure 3 – Word Embedding Visualization*

In Figure 3, the spatial arrangement of words indicates their semantic relationships. The clustering of most data points suggests that a significant portion of words share similar meanings or contexts. However, there are a few outliers scattered across the plot, representing rare words or terms with unique semantic attributes. This distribution pattern underscores the richness and diversity of language use within the dataset, with common themes dominating while rare or specialized terms exist as exceptions.

**Discussion:**

By utilizing bigrams as features for the Naive Bayes classifier, we observed a significant increase in accuracy, achieving an overall accuracy rate of 65%. This observation is consistent with the research of *Carrero et al.,*[8], who similarly identified bigrams and trigrams as the most effective features for classification, outperforming unigrams. Another study that supports our findings is by *Wicentowski et al.,*[9], who found that a basic NB model trained on word bigrams performs comparably to expert human annotators, but less effectively when smoking cues are absent. In the study done by *Uzuner et al.,*[10], 12 systems were run with micro-averaged F-measures above 0.84 utilizing different classifiers.

In our study focusing on Named Entity Recognition (NER) and relation extraction, we observed significant successes alongside notable challenges. The effective extraction of entities like "alkaline phosphatase" and "bone" from medical text data underscores the viability of our methodology. However, the inability to establish relations between these entities highlights the intricate nature of syntactic structures and contextual nuances present in medical records, posing challenges to accurate extraction. This underscores the necessity for further refinement in relation extraction techniques to precisely capture semantic associations between entities. Moreover, while our system demonstrated satisfactory performance in ALP value extraction, encountered challenges such as formatting inconsistencies and numerical representation variations underscore the significance of preprocessing techniques for enhancing extraction accuracy.

Top of Form

**Conclusion:**

This study underscores the potential of NLP in augmenting clinical analysis by extracting actionable insights from unstructured medical text data. While advancements in smoking status classification demonstrate promising results, further research is warranted to address challenges in relation extraction, particularly regarding ALP and bone associations. By leveraging NLP techniques, healthcare professionals can unlock new avenues for understanding patient data, conducting medical research, and ultimately improving clinical decision-making processes in the era of big data analytics.

**References:**

1. Koleck, T. A., Dreisbach, C., Bourne, P. E., & Bakken, S.

(2019). Natural language processing of symptoms

documented in free-text narratives of electronic health

records: a systematic review. *Journal of the American*

*Medical Informatics association*, *26*(4), 364-379.

<https://doi.org/10.1093/jamia/ocy173>

1. Sheikhalishahi, S., Miotto, R., Dudley, J. T., Lavelli, A.,

Rinaldi, F., & Osmani, V. (2019). Natural language

processing of clinical notes on chronic diseases:

systematic review. *JMIR medical informatics*, *7*(2),

e12239. <https://doi.org/10.2196/12239>

1. Hao, T., Huang, Z., Liang, L., Weng, H., & Tang, B.

(2021). Health Natural Language Processing:

Methodology Development and Applications. JMIR

medical informatics, 9(10), e23898.

<https://doi.org/10.2196/23898>

1. Ong, K. C., Cheong, G. N., Prabhakaran, L., & Earnest, A.

(2005). Predictors of success in smoking cessation

among hospitalized patients. *Respirology*, *10*(1), 63-69.

<https://doi.org/10.1111/j.1440-1843.2005.00656.x>

1. Rajendran, S., & Topaloglu, U. (2020). Extracting

Smoking Status from Electronic Health Records Using

NLP and Deep Learning. *AMIA Joint Summits on*

*Translational Science proceedings. AMIA Joint*

*Summits on Translational Science*, *2020*, 507–516.

1. Chen, H., Li, J., & Wang, Q. (2018). Associations between

bone-alkaline phosphatase and bone mineral density in

adults with and without diabetes*. Medicine, 97(17),*

*e0432.* DOI: 10.1097/MD.0000000000010432

1. n2c2 NLP Research Data Sets. (n.d.).

Portal.dbmi.hms.harvard.edu.

<https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/>

1. Carrero, F., Hidalgo, J. G., Puertas, E., Maña, M., & Mata,

J. (2006). Quick prototyping of high performance text

classifiers. In *i2b2 Workshop on Challenges in Natural*

*Language Processing for Clinical Data*.

1. Wicentowski, R., & Sydes, M. R. (2008). Using implicit

information to identify smoking status in smoke-blind

medical discharge summaries. *Journal of the American*

*Medical Informatics Association*, *15*(1), 29-31.

<https://doi.org/10.1197/jamia.M2440>

1. Uzuner, O., Goldstein, I., Luo, Y., & Kohane, I. (2008).

Identifying patient smoking status from medical

discharge records. *Journal of the American Medical*

*Informatics Association: JAMIA*, *15*(1), 14–24.

<https://doi.org/10.1197/jamia.M2408>

‌