# Applying Text Classification Techniques to Locally Generated Clinical Notes

**Andrew K. Teng, MS[1], Adam B. Wilcox, PhD[1]**
**[1]University of Washington, Biomedical Informatics and Medical Education**

## Introduction

Since the introduction of electronic health records, large amounts of data in the clinical environment are being recorded. Often times, the data can be structured, such as vital signs; however, many times the data can be unstructured, such as clinical notes, and these data are often significantly more difficult to analyze. However, these texts often times contain meta information that may not explicit within the structured data and can create new domains for analysis. We collaborated with a local hospital that has a housing unstable patient population of about 6.5% [1]. Additionally, housing instability can often result in higher uses of acute care [2]. Although there have been previous attempts [3], our approach differs as we utilize open source Python packages to test simple text classification methods that can easily generalizable and implemented.

## Methods and Results

We first extracted structured data from acute care patients over a one year timeframe. Once we gathered a patient list, we then extracted clinical text data (including social history, habits, and opioid use) over the past five years for this list of patients. To lower the scope of our exploration, we decided to look at social factors, specifically housing stability. We queried the clinical text from various sources, such as Admit and ED notes, and extracted chunks of text that were related to social history. To verify that we were extracting the social history correctly, we verified by manual chart review a random set of ten patients. Once confirmed, we extracted the clinical text and manually labelled the sentiment of the text, treating each entry independently. We extracted 21,876 had social history entries, of which 2,408 were manually reviewed. Due to missing data, only 1,785 rows were manually labelled as “housing stable” (0) and “housing unstable” (1), covering 200 unique patients, of which 71 (35%) are scored homeless, and 1,361 unique encounters. Three different models with two different feature selection methods (bag of words (BOW) and bigrams) were used to classify and predict housing stability for the extracted clinical text.

| **Model** | **Accuracy** | **Precision (0/1)*** | **Recall (0/1)*** | **F1 score (0/1)*** |
|---|---|---|---|---|
| BOW naïve Bayes | 0.901 | 0.90/0.90 | 0.93/0.87 | 0.91/0.88 |
| BOW stochastic gradient descent | 0.927 | 0.93/0.93 | 0.95/ 0.90 | 0.94/0.91 |
| BOW logistic regression | 0.873 | 0.87/0.88 | 0.92/0.82 | 0.89/ 0.85 |
| Bigram naïve Bayes | 0.892 | 0.88/0.92 | 0.94/0.82 | 0.91/0.87 |
| Bigram stochastic gradient descent | 0.922 | 0.92/0.92 | 0.94/0.89 | 0.93/0.91 |
| Bigram logistic regression | 0.912 | 0.91/0.91 | 0.93/0.88 | 0.92/0.90 |

## Discussion and Conclusion

From our preliminary analysis, we can see that there is slight variation in the accuracy amongst text classifiers. However, we can see that there is high accuracy and high scoring metrics in terms of classifying housing stability from unstructured clinical text notes. There were many limitations to our preliminary analysis: (1) Often times when social factors were unable to be assessed due to patient condition, notes were either left empty or a previously recorded entry was copied forward, yielding in 5.7% of the social history text as duplicates, (2) housing stability notes were complex as homelessness qualities were recorded differently amongst providers (e.g. shorthand and spelling of local facilities), and (3) patients can falsify or downplay their living situation and other social factors due to embarrassment. For future work, we intend on expanding our extraction and text classification to a wider range of other social factors. Furthermore, we are considering the development of a local lexicon dictionary or parser to include specific and locally dependent variables, such as local housing structures, programs, and shelters.

## References

1. Clemenzi-Allen A, Neuhaus J, Geng E, et al. Housing Instability Results in Increased Acute Care Utilization in an Urban HIV Clinic Cohort. *Open Forum Infectious Diseases*. 2019;6(5).
2. Boonyaratanakornkit J, Ekici S, Magaret A, et al. Respiratory Syncytial Virus Infection in Homeless Populations, Washington, USA. *Emerging Infectious Diseases*. 2019;25(7):1408-1411.
3. Gundlapalli AV, Carter ME, Palmer M, et al. Using natural language processing on the free text of clinical documents to screen for evidence of homelessness among US veterans. *AMIA Annu Symp Proc*. 2013;2013:537-546.