Classifying Laboratory Test Results Using Machine Learning

William Lu, Kenny Chiu, Joy (Sizhe) Chen, Nilgoon Zarei

November 16th, 2018

1 Introduction

An important service provided by public health organizations worldwide is the analysis of population-level disease trends [1] [2]. These analyses aid efforts for proactive prevention of communicable diseases and provide insight into the historical patterns of pathogen spread. Population-level disease surveillance is also utilized by public health authorities to identify outbreaks and diagnose and treat illnesses [2].

Disease surveillance efforts are heavily data-driven, requiring the extraction of structured information from unstructured or semi-structured data. The volume of data is often extremely large; for example, a previous effort by the British Columbia Centre for Disease Control (BCCDC) to monitor patients infected with Hepatitis C required the manual preparation of an anonymized database containing the health records of 1.5 million patients [3]. The Centers for Disease Control and Prevention (CDC) receives approximately 20 million laboratory reports annually [2]; the BCCDC's data warehouse contains over 330 million test results spanning over 20 years.

Increasingly, semi-structured text is the dominant format of clinical data and lab result descriptions, with typical health documents consisting of approximately 60% structured information and 40% free-form text [4] [5]. The task of extracting structured information from semi-structured lab result descriptions is currently being performed manually by domain experts. Due to the large volume of data, this manual processing is expensive to carry out, slow, and error-prone.

In this paper, we present a machine learning approach that has demonstrated success in automating this process, achieving human-level ($> 95\%$) accuracy. We work with a subset of the lab result descriptions from the BCCDC's data warehouse, comprised of semi-structured text data. Our approach extracts four structured labels: a binary *Test Performed* label describing whether the test was performed; a 4-class *Test Outcome* label describing whether the test result is positive, negative, indeterminate, or missing; and multi-class *Organism Genus* and *Organism Species* labels listing the positive organisms in the test.

We begin our paper with a review of previous literature relating to similar data mining tasks. We follow with a description of the specifics of the dataset we used to benchmark the performance of our machine learning classifiers. We then present our machine learning models alongside an evaluation of the results achieved on our validation dataset. We conclude with a summary of the limitations of our research, including remarks about the generalizability of our models to novel datasets.

2 Related Work

In 2018, Segura-Bedmar et al. evaluated the performance of standard machine learning classifiers in classifying electronic medical records as either positive (describing a case of anaphylaxis) or negative (not describing such a case) [6]. The classifiers they tested included Multinomial Naive Bayes, Logistic Regression, Random Forest, and Linear Support Vector Machine (SVM). The dataset used in this work exhibits class imbalance: less than 1% of the records are labelled positive. This class ratio is similar to the class ratio of the binary *Test Performed* class in our dataset. A modified k-means algorithm was used to downsample the majority class by replacing similar instances with their cluster centroid. The researchers found that Logistic Regression and Linear SVM achieved around 97% precision and 93% recall. All the records used to evaluate the classifiers are written in Spanish, unlike our dataset which consists of English text. We expect that the differences in linguistic structure and semantics will be significant enough to produce differing results.

In 2014, Velupillai et al. presented a symbolic assertion-based classifier for detecting whether a disorder described in a snippet of semi-structured clinical text is affirmed, negated, or uncertain [7]. This is similar to our problem of classifying the *Test Outcome* of a lab result as positive, negative, or indeterminate, or missing; however, it differs due to the lack of a “missing” class. The researchers specifically aim to construct a system capable of accurately determining the scope and interpretation of negation words such as “not” that appear in the clinical text. Their final natural language processing pipeline achieved an overall F-score of 83% on a corpus of clinical text written in Swedish; again, we expect that the differences between English and Swedish will significantly affect the results of text mining.

Jang et al. used hidden Markov models to text mine doctors’ notes in 2006 [8]. The document corpus they worked with was more challenging as the notes were written in a mixture of English and Korean. Their models achieved around 60%-70% accuracy and aimed to be robust to unknown phrases not seen in the training corpus. One notable tool they also used is *MetaMap*, which annotates input text with medical tags and semantics. MetaMap is also commonly used in other health domains, such as text mining for cancer-related information [9].

In 2014, Kang and Kayaalp explored the problem of extracting laboratory test information from biomedical text [10]. They compared the performance of an original symbolic information extraction system to various machine learning-based NLP systems. Their results showed that well-tailored symbolic approaches may outperform machine learning-based approaches. Kang and Kayaalp used a collection of decision summaries from the U.S. Food and Drug Administration as their document corpus. These summaries are written in natural language, unlike our test result descriptions which are written in point-form. We expect this difference will be significant enough that a symbolic approach to our problem will require more complex logic before it can achieve similar results to theirs.

Hasan et al. evaluated the performance of standard machine learning classifiers for multi-class annotation of text data in their 2016 paper [11]. The dataset they used was a collection of around 11,000 transcripts of clinical interviews between psychologists and adolescents or caregivers. They found that a Support Vector Machine classifier was able to annotate conversational snippets with codes from a 17-class codebook with 70.3% accuracy, although the accuracy decreased to 53.7% when a 41-class codebook was used. This problem is similar to our problem of extracting organism names from lab result descriptions, due to the large number of classes in both problems and the semi-structured, grammatically incorrect form of the data. However, the transcripts used by Hasan et al. do not contain abbreviations such as "hbsag" and lab terminology such as "PCR", both of which are prominent features of our dataset.

In a 2016 study [12], Napolitano, Marshall, Hamilton, and Gavin used the k-nearest neighbours (KNN) machine learning algorithm to text mine surgical pathology reports. Their approach was able to identify chunks of text in the reports that provide information about the morphology of the cancer tumours detected and the final diagnoses. Their document corpus is a collection of semi-structured and unstructured reports from the Northern Ireland Cancer Registry. Similarly to our dataset, the semi-structured nature of the reports comes from the ability of the clerical staff to automatically generate report templates and freely edit them or append additional free text. However, the reports are long, multi-paragraph documents written in complete sentences, as opposed to the short phrases that comprise the lab result descriptions in our dataset. The researchers found that the KNN classifier achieved precision and recall scores of above 90% on the semi-structured reports, and also indicated that classifiers trained on semi-structured data do not generalize well when predicting on completely unstructured data.

3 Dataset

Our dataset is a subset of the data available in BCCDC's data warehouse, and consists of approximately 950,000 test results. The dataset contains no personally identifying patient information. Some test results in the dataset are results of proficiency tests, which are routine sanity checks run by the BCCDC to make sure their lab equipment is working properly. Some other lab reports in the dataset consist purely of a floating-point number such as "0.016". All proficiency tests and purely numeric results were filtered from the dataset before further analysis was performed, which reduced the size of the dataset to approximately 360,000 test results.

Each test result consists of the original laboratory report, stored as semi-structured text, along with structured metadata stored in additional database columns. The metadata accompanying the lab reports codifies:

- The test code
- The test type (culture, antibody, NAT/PCR, etc.)

The semi-structured form of the lab reports is the consequence of the laboratory technician having the option of either inputting a code that auto-generates the text, inputting completely free-form text, or appending custom free-form text to the auto-generated text. In addition, As a result, there is no consistent structure in the test results that allows for simple rule-based processing. All the text is written in English, but many results are written in short phrases as opposed to full sentences, and do not conform to standard English syntactical and grammatical rules. Examples of the semi-structured lab reports are listed below:

> *Specimen rejected | Test not performed. | No evidence of HCV infection.*
>
> *No Bordetella pertussis DNA detected by PCR.*
>
> *Result inconclusive. | Culture results to follow. | Varicella Zoster Virus | 'Isolated.'*
>
> *'Organism identified as:' | Haemophilus influenzae | Biotype | | non serotypable (non encapsulated)*

Some test results contain spelling, grammatical, and typographical errors. For example, "serotype" is misspelled as "seretype" and "group" is misspelled as "froup" in the result text below:

> *BCCDC seretype: non froup 5 | Final | 12/Jun/2009 | Sputum | Streptococcus pneumoniae | STUDY*

The word 'not' is often used ambiguously, as it could be interpreted as negating the entire test result, negating the current observation, or as part of a subtype of an organism, among many other interpretations:

> *NEGATIVE for Shiga toxin stx1 and stx2 genes by PCR. | Isolate serotyped as: | Escherichia coli | not | O157:H7*
>
> *Isolate not | Salmonella species*

Some test results contain many negative organism names. For example, the following lab result text contains all organisms that multiplex NAT is capable of detecting. However, only Rhinovirus, Enterovirus, and Adenovirus were positive in the actual test:

> *Rhinovirus or Enterovirus detected by multiplex NAT. | | Adenovirus detected by multiplex NAT. | | Multiplex NAT is capable of detecting Influenza A and B, Respiratory Syncytial Virus, Parainfluenza 1, 2, 3, and 4, Rhinovirus, Enterovirus, Adenovirus, Coronaviruses HKU1, NL63, OC43, and 229E, hMetapneumovirus, Bocavirus, C. pneumoniae, L. pneumophila, and M. pneumoniae. | | MULTIPLE INFECTION DETECTED*

Due to clinical procedures, early observations recorded in the lab reports that were invalidated by later tests cannot be erased. Thus, some lab reports contain contradictory information. For example, only Moraxella osloensis is positive in the result description below, despite the text mentioning Neisseria meningitidis as well:

> *Organism identified as: | Neisseria meningitidis nongroupable | Upon further investigation | Organism identified as: | Moraxella osloensis | by 16S rRNA gene sequence analysis.*

The most up-to-date information is not always written at the end of the result description. For example, the test corresponding to the result text below was not performed, despite the last observation stating the contrary:

> *TEST NOT PERFORMED | Galactomannan testing is valid only for Haematology and lung transplant patients with no recent antifungal exposure | Test performed at Provincial Laboratory of Public Health, Edmonton*

Dates?

4 Approach

5 Results

6 Discussion and Future Work

7 Conclusion

References

[1] BC Centre for Disease Control. (n.d.). Retrieved from http://www.bccdc.ca.

[2] Surveillance Strategy Report - Electronic Laboratory Reporting. (n.d.). Retrieved from https://www.cdc.gov/surveillance/initiatives/lab-reporting.html.

[3] BCCDC publishes new findings on power of integrated health data. (2016, September 8). Retrieved from http://www.bccdc.ca/about/news-stories/news-releases/2016/integrated-health-data.

[4] Suominen, H. (2014). Text mining and information analysis of health documents. *Artificial Intelligence in Medicine, 61*(3), 127-130.

[5] Dalianis, H., Hassel, M., & Velupillai, S. (2009). The Stockholm EPR Corpus - Characteristics and Some Initial Findings. *Proceedings of the 14th International Symposium on Health Information Management Research*, 14-16.

[6] Segura-Bedmar, I., Colón-Ruíz, C., Tejedor-Alonso, M., & Moro-Moro, M. (2018). Predicting of anaphylaxis in big data EMR by exploring machine learning approaches. *Journal of Biomedical Informatics, 87*(November 2018), 50-59.

[7] Velupillai, S., Skeppstedt, M., Kvist, mowery, D., Chapman, B. E., Dalianis, H., Chapman, W. W. (2014). Cue-based assertion classification for Swedish clinical text - Developing a lexicon for pyConTextSwe. *Artificial Intelligence in Medicine, 61*(3), 137-144.

[8] Jang, H., Song, S.K., & Myaeng, S.H. (2006). Text mining for medical documents using a hidden Markov model. *Lecture Notes in Computer Science* (Vol. 4182, pp. 553-559).

[9] Spasic, I., Livsey, J., Keane, J. A., & Nenadic, G. (2014). Text mining of cancer-related information: Review of current status and future directions. *International Journal of Medical Informatics, 83*(9), 603-623.

[10] Kang, Y. S. & Kayaalp, M. (2013). Extracting laboratory test information from biomedical text. *Journal of Pathology Informatics, 4*(1), 4-23.

[11] Hasan, M., Kotov, A., Carcone, A. I., Dong, M., Naar, S., & Hartlieb, K. B. (2016). A study of the effectiveness of machine learning methods for classification of clinical interview fragments into a large number of categories. *Journal of Biomedical Informatics, 62*, 21-31.

[12] Napolitano, G., Marshall, A., Hamilton, P., & Gavin, A. T. (2016). Machine learning classification of surgical pathology reports and chunk recognition for information extraction noise reduction. *Artificial Intelligence in Medicine, 70*, 77-83.