ORIGINAL RESEARCH PAPER

Computer Science

SENTIMENT ANALYSIS AND MACHINE LEARNING ON CLINICAL TEXT: AN OVERVIEW

KEY WORDS: Sentiment Analysis, Machine Learning, NLP, Text Mining

Mily Lal Assistant Professor, DYPIEMR, Akurdi Pune 411044.

Mohini Avatade* Assistant Professor, DYPIEMR, Akurdi Pune 411044. *Corresponding Author

Rachana Mudholkar Assistant Professor, DYPIEMR, Akurdi Pune 411044.

ABSTRACT

This paper presents an overview of the various semantic analysis methods on clinical text and with the support of machine learning techniques will help in identification of documentation reflecting advanced health care planning measures and serious illness symptoms. Doctors document their communication as unstructured free text in clinical notes. Routine assessment of unstructured data is time consuming and costly. Semantic analysis methods can offer an efficient way to access and predict within electronic health record data, helping to better quality of care.

INTRODUCTION

To improve patients care, medical practitioners must communicate with them about their treatment preferences. In the absence of explicit care decisions, medical practitioners may provide medical care [1] that does not provide a significant benefit to the patient. Therefore it is recommended that care preferences are discussed and documented in the EHR (eHealth records). Appropriate utilization of EHR databases could have an immersive impact on health care research and delivery. Owing to the large amount of unstructured data available in EHRs, there has been a concomitant increase in research to Semantic Analysis methods and applications for the clinical domain [2], [3]. EHR-derived data offers the most immediate value, given the time and patient numbers over which data have already been collected, their comprehensive and now-established use across healthcare, and given the depth of clinical information potentially available in real world services. [4]

The role of machine learning in natural language processing (NLP) and Semantic Analysis is to improve, accelerate and automate the underlying text analytics functions that turn unstructured text into useable data and insights. [5] Another motivation for this paper is to gain a concrete understanding of the under-utilization of Semantic Analysis and Machine Learning in EHR-based clinical research.

CHALLENGES

In order to ensure that performance of the system to be appropriate the application involving NLP and Semantic Analysis must undergo evaluation. The system must have adequate recall, precision, and specificity for the intended application and it should be possible to adjust its performance according to the needs of the application. [6]

An enormous collections of online patient records in textual form will be available to the developers which are confidential. All personal identifying information must be detected and removed to comply with laws protecting patient confidentiality. [6]

The system has to be able to function well in heterogeneous health care facilities and for different clinical applications. Therefore the system will have to handle different interchange formats (i.e. XML, HL7) and heterogeneous formats. The system will also have to generate output that can be stored in an existing repository, but the output often has complex and nested relations, and it may be impossible to map the output without substantial loss of information. Besides these, the output has to be comparable so that it can be used across institutions for a variety of automated applications. The system should contain a controlled vocabulary and a standard representation for the domain. [6]. Another major challenge faced is the evaluation of the system as it is difficult to share data among different institutions. Obtaining a standard for evaluation is very time consuming and costly. [6]

The language used in the clinical text can be extremely expressive .There are often different ways to describe the same medical concept and also numerous ways to express modifiers of the concept. Such modifiers make it more complex to retrieve reports, similarly the query for retrieving the information will have a very fine granularity. [6]

There is no standardized structure for the clinical reports. The format for text within the reports is not standardized. Punctuation is often missing or is inappropriate, and a new line may be used instead of a period to signify the end of a sentence. Some reports contain tables with different configurations as well as text. [6]

The clinical reports may contain abbreviations that they are highly ambiguous. An additional problem is that a unique abbreviation may be defined in a single report. An automated system would have to automatically capture the implicit information based on domain knowledge. For example, when mass occurs in a radiological report of the chest, it means mass in lung whereas if it occurs in a mammography report, it means mass in breast. [6]

Domain knowledge along with knowledge of the report structure is mandatory in order to associate findings with possible diagnoses. Machine learning techniques can be used can be used to develop the components of the system based on complex rules. Semantic Analytics generally need a large number of training examples to train, refine, or test the system. Rare events may be difficult to find a large number of reports for these events. [6]

COMPONENTS OF THE SYSTEM

As the first step the clinical report will be pre-processed in order to identify segments and to handle textual irregularities. This step follows a NLP processor and Semantic Analyzer. It uses domain knowledge associated with syntactic and semantic features, and then is stored in a clinical EHR database. Once the data is structured, it can be used for different clinical applications. The extracted EHR information for an application typically requires additional medical knowledge and machine learning techniques to identify advanced care planning measures, serious illnesses, and serious illness symptoms.

In order to find the structural units of information and their relations of the sentences, POS Tagger is frequently used. POS Tagger incorporates syntactic knowledge and contextual features of words surrounding a syntactically ambiguous word to determine the most likely syntactic category.[5]

A large domain corpus that has been syntactically tagged is required to be used as a training corpus to establish the appropriate parse probabilities. A partial syntactic analysis is a much simpler, more efficient, and more robust process, but has lower precision. It could be used to detect noun phrases.[8] Partial analysis is not adequate for clinical applications, such as an alerting application.

Semantic analyzer is used to semantically classify words and phrases that occur in the domain, and then establishes an interpretation of their relationships. Unsupervised Machine Learning techniques like Latent Semantic Indexing (LSI) and Clustering can be used to identify semantic relationships among words and phrases. LSI identifies on words and phrases that frequently occur with each other whereas Clustering groups similar documents together into groups or sets.

These groups are the sorted based on relevancy score.

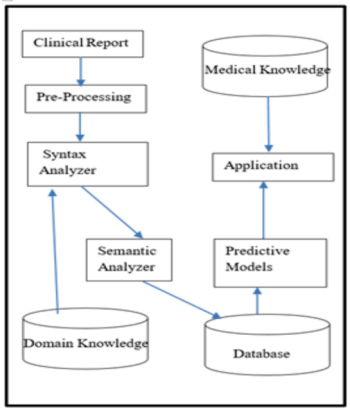
Supervised Machine Learning Techniques like Support Vector Machines and Deep learning can be used to build and improve the semantic analytics functions and NLP features.

Many efforts are taken develop user-centric tools for machine learning and NLP making it easier for the end users to use them.[9] A closed loop system can be created that can be used to build continuously improving predictive models by learning iteratively from their end-users.[10],[11],[12].

Predictive models based on Logistic Regression, Neural Networks, Decision Trees and Nearest Neighbors has been proposed to with the aim to predict patient's state in certain periods of time that can vary from very short time prediction, like respiratory issues, heart diseases or the need of intense care; to med-long term predictions such as future diseases or life expectancy[13]. Neural networks are often compared considered better than Logical regression, to predict categorical data, as cancer diagnostic and other diseases.[14] It also can be used to predict numerical data, as hospital stay times of patients.[15]. While Decision Trees are useful for real time analysis since the resulting models can make fast predictions.[16]

K-Nearest Neighbor can be used to predict numerical or categorical results. K-Nearest Neighbor works better when the data has well defined clusters, since predictions are based on distances between data.[17].

Figure 1: Components of an Application in the Clinical Domain



CONCLUSIONS

NLP for the clinical domain has become an important part of advancing data-driven health care research. It involves the integration of many forms of knowledge, including syntactic, semantic, lexical, and domain knowledge. The main challenges in using NLP in clinical data include availability of data, confidentiality, evaluation and reporting standards of the data. Machine Learning techniques can help the industry to analyze the ever-increasing complexity of today's clinical data and manage continuous data collection from emerging data resources. Machine learning is a powerful approach to predictive modeling. Therefore, NLP with the support of machine learning techniques will help in identification of documentation reflecting advanced health care planning measures and illness symptoms.

REFERENCES:

[1] Thanh Huynh, Prince Raj, Eric Kleerup, and Neil Wenger. The opportunity cost of futile treatment in the intensive care unit. Critical Care Medicine, 42(9):1922–82, September 2014
[2] A. Névéol, P. Zweigenbaum Clinical Natural Language Processing in 2014: foundational methods supporting efficient healthcare Yearb. Med. Inform., 10 (1) (2015), pp. 194-198.
[3] S. Velupillai, D. Mowery, B.R. South, M. Kvist, H. Dalianis Recent advances in clinical natural language processing in support of semantic analysis IMIA Yearb. Med. Inform., 10 (2015), pp. 183-193
[4] J. M. Rabaey, Digtal Integrated Circuits— A Design Perspective Upper Saddle River, NJ: Prentice-Hall, 2001
[5] https://www.lexalytics.com/lexablog/machine-learning-vs-natural-language-processing-part-1
[6] Carol Friedman, "Semantic Text Parsing For Patient Records," Chapter 15, Department of Biomedical Informatics, Columbia University, New York, New York 10027.
[7] https://blogs.oracle.com/health-sciences/how-machine-learning-is-transforming-clinical-data
[8] Hersh, W., Mailhot, M., Arnott-Smith, C., and Lowe, H. (2001). "Selective Automated Indexing of Findings and Diagnoses in Radiology Reports," J.Biomed.Inform.; 34(4):262273
[9] D'Avolio, L. W., Nguyen, T. M., Goryachev, S., and Fiore, L. D. Automated concept-level information extraction to reduce the need for custom software and rules development. Journal of the American Medical Informatics Association 18, 5 (2011), 607–613.
[10] Amershi, S., Lee, B., Kapoor, A., Mahajan, R., and Christian, B. Cuet: Human-guided fast and accurate network alarm triage. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '11, ACM (New York, NY, USA, 2011), 157–166.
[11] Heimerl, F., Koch, S., Bosch, H., and Ertl, T. Visual classi er training for text document retrieval. Visualization and Computer Graphics, IEEE Transactions on 18, 12 (2012), 2839–2848
[12] Fails, J., and Olsen Jr, D. Interactive machine learning. In Proceedings of the 8th international conference on Intelligent User Interfaces, ACM (2003), 39–45
[13] Herland, Matthew, Taghi M. Khoshgoftaar, and Randall Wald. "A review of data mining using big data in health informatics." Journal of Big Data 1.1 (2014): 1-35.
[14] Khan, Javed, et al. "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks." Nature medicine 7.6 (2001): 673-679
[15] Walsh, Paul, et al. "An artificial neural network ensemble to predict disposition and length of stay in children presenting with bronchiolitis." European Journal of Emergency Medicine 11.5 (2004): 259-264.
[16] Zhang, Yang, et al. "Real-time clinical decision support system with data stream mining." BioMed Research International 2012 (2012)
[17] A. Salcedo-Bernal, M.P. Villamil-Giraldo, A.D. Moreno-Barbosa. "Clinical Data Analysis: An opportunity to compare machine learning methods." Conference on Enterprise Information Systems / International Conference on Project Management / Conference on Health and Social Care Information Systems and Technologies, CENTERIS / Projman / HCist 2016, October 5-7, 2016