International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

A Novel VTE Prediction Model using Natural Language Processing (NLP) and Machine Learning Methods

Nayan Pawar, Dr. (Prof). Amol Potgantwar and Dr. (Prof).Mangesh Ghonge

Department of computer Engineering Sandip institute of technology and research centre Nashik,India

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

Padua straight model is generally utilized for the hazard appraisal of venous thromboembolism (VTE), which is a typical and preventable confusion for inpatients. Separating VTE chance components from unstructured medicinal records in emergency clinic can comprehend VTE occasions and create productive hazard appraisal model. In this investigation, we proposed a philosophy based technique to mine VTE hazard factors joining natural language processing (NLP) and AI (ML) strategies. Medici- nal records of 3106 inpatients were prepared and terms in different ontologies from different segments of records improved in VTE patients were arranged consequently. At that point ML strategies were utilized to assess terms' significance and terms in- side conceding analysis and progress notes indicated preferred VTE forecast execution over different segments. At last a novel VTE forecast model was assembled de- pendent on chose terms and demonstrated higher AUC score (0.815) than the Padua model (0.789).

Keywords: *Medical Record, Venous thromboembolism (VTE), Natural Language Processing (NLP), Risk Assessment, Machine Learning (ML).*

Introduction

In this modern age, 1 in 4 people worldwide are dying from conditions caused by VTE(Venous Thromboembolism) . And it's a starting fact that up to 900,000 people in the United States alone are affected by blood clots each year. American guide- lines has reported that hospital admission is related to nearly 25% of all VTE patients and 10% deaths of inpatient is caused by PE. As a typical complexity for inpatients, venous thromboembolism (VTE) containing pulmonary embolism (PE) and deep venous thrombosis (DVT) is a preventable reason for death. Since it is firmly iden- tified with ethnic foundation and sickness range, Study varied from the Caucasians in the illness hazard appraisal, which caused horrible showings of the Padua model suggested by American College of Chest Physicians when applied inpatients in the Internal Department. Along these lines, to locate the potential VTE chance factor and create forecast model determined to inpatients are justified. In addition, the quick advancement of therapeutic informatization and electronic health record (EHR) framework permits the collection of expanding number of medicinal records and gives the probability of researching ailments in increasingly intricate and exact strategies, contrasted and conventional methodologies with little example size and less factors. Numerous analysts have contemplated connections between different ailments and hazard factors utilizing Machine Learning (ML) and Natural Language Processing (NLP) techniques and demonstrated promising outcomes. thought about prescient validities of numerous ML strategies on cardiovascular hazard appraisal examined the Alzheimer's sickness hazard by regularized calculated relapse, prepared the help vector machine to do VTE chance expectation for malignant growth patients. Nonetheless, over investigations' highlights are pre-planned and constrained, which can scarcely take full favorable circumstances of amounts of patient data in medicinal records and find new information, for example, other potential factors related with the illness. Some profound learning models [5, 6] are additionally proposed to consolidate medicinal ontologies to dissect high-dimensional and heterogeneous restorative records yet their outcomes absence of interpretability. So as to enable the clinicians to investigate new competitor VTE hazard factors and create proficient expectation model with certain interpretability from medicinal records, we propose an ontology-based approach which processes the free-text in medical records carefully, evaluate significance of terms from ontologies consequently lastly builds the model dependent on applicant terms. The entire work process needs no clinicians direction and the created outcomes can rouse further medicinal investigations. II.

Literature Survey

[1] In March 2017, Stephen F. Weng, Jenna Reps, Joe Kai, Jonathan M. Garibaldi, Nadeem Qureshi studied relationship between various diseases and risk factors using machine learning (ML) and natural language processing (NLP) methods and showed Machine-learning significantly improves accuracy of cardiovascular risk prediction, increasing the number of patients identified who could benefit from preventive treatment, while avoiding unnecessary treatment of others.
[2] In 2016, Edward Choi 1 , Mohammad Taha Bahadori 1 , Le Song 1 , Walter F. Stewart 2 , Jimeng Sun introduce GRAM: Graph-based Attention Model for Healthcare Representation Learning proposed GRAM, a graph-based attention model using both a knowledge DAG and EHR to learn an accurate and interpretable representations for medical concepts. [3] In 2016, Riccardo, M., Li, L., Kidd, B.A., and Dudley, J.T. introduce 'Deep Pa- tient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records' Produce a system which examine combining comprehensive terms don't achieve the best prediction result. [4] In 2013, Ramon Casanova, Fang-Chi Hsu, Kaycee M. Sink, Stephen R. Rapp, Jeff D. Williamson, Susan M. Resnick, Mark A. Espeland, for the Alzheimer's Disease Neuroimaging Initiative introduce Alzheimer's Disease Risk Assess- ment Using Large-Scale Machine Learning Methods. In this compared pre- dictive validities of multiple ML methods on cardiovascular risk assessment, Casanova, et al. [5] In 2013, Ko¨hler, S., Doelken, S.C., Mungall, C.J., Bauer, S., Firth, H.V., Bailleul-Forestier, I., Black, G.C., Brown, D.L., Brudno, M., and Campbell, J. Introduce 'The Human Phenotype Ontology project: linking molecular biol- ogy and disease through phenotype data', Nucleic acids research Explain the Human Phenotype Ontology. [6] In march 2010, S. Barbar, f. Noventa, v. Rossetto, a. Ferrari, b. Brandolin, m. Perlati, e.debon , d. Tormene, a . Pagnan and P.Prandon introduced 'A risk assessment model for the identification of hospitalized medical patients at risk for venous thromboembolism: the Padua Prediction Score'. In this explain the Padua Prediction Score and poor performances of the Padua model on patients other than Western patients.

Proposed Methodology

In order to help the clinicians explore new candidate VTE risk factors and develop efficient prediction model with certain interpretability from medical records, we pro- pose an ontology-based approach which processes the free-text in medical records carefully, evaluates importance of terms from ontologies automatically and finally constructs the model based on candidate terms. The whole workflow needs no clinicians guidance and the generated results can inspire further medical studies. The overall dissertation is divided into following parts: 1. Medical Records 2. Ontology Sources 3. Ruled-Based Section Extraction 4. Automatic Ontology Enrichment 5. Ontology And Section Evaluation By ML Methods Medical Records: Medical records of inpatients need to collect as a dataset and every patient had two documents, admission note and progress note, which were both unstructured and had lots of paragraphs consisting of free text. patient condition. Ontology Sources: In order to obtain comprehensive information of patients involving symptoms, diagnoses, drugs, operations and so on, multiple kinds of ontologies were gathered. Rule-Based Section Extraction: The workflow of ontology extraction and risk assessment model establishment was shown in Fig. of Architectural Flow of System Automatic Ontology Enrichment: With sections in medical records, we then searched terms in ontologies existed in sections and sorted them according to their characteristics and weights in documents to shrink the size of term set. Ontology And Section Evaluation By Ml Methods: Next the importance of terms to VTE risk assessment was evaluated in unit of the section. Features of specific section were constructed based on terms within this section. A. Architecture The architectural of ontology mining and VTE prediction model construction from medical records Shown in Fig Architecture Diagram. Different algorithms and methods used for implementing the system. As shown in fig. Architecture Diagram, the algorithm will perform following steps: 1) The Admit note and progress not got as medical records. The admit note usually included 11 sections: chief complaint, present history, previous history, personal history, family history, obsterical history, menstrual history, physical examination, lab- oratory examination, admitting diagnosis and physician's signature, and the progress note had daily description about patient condition. 2) Ontology Sources like (MeSH),Human Phenotype Ontology (HPO)[5], SNOMED and ICD-10. After preprocessing, 37111 ICD codes, 11903 phenotypic abnormal- ities from HPO, 55750 MeSH words, and 11652 terms from SNOMED were saved and merged as the final ontology set. 3)

The first step was parsing sections within medical records. Due to the pattern that sections started with specific titles and white space existed between two sections, we utilized the regular expression to capture the start position of one section, and considering that some sections were removed by clinicians for convenience and orders of them could be changed, a greedy match algorithm was implemented to find the title of section having the minimum distance with current position iteratively. The text between two titles was regarded as a section related to the first title and current position in the document was updated after a section was recognized. One example of the algorithm was shown in Fig. of Example of the Greedy Algorithm For the admit note, 8 sections were parsed excluding the obsterical history, menstrual history and physician's signature. For the progress note, the daily objective description were split via the date.

In a medical records, The ontologies list term existed in sections and sorted them according to their characteristics and weights in documents to shrink the size of term set. For every kind of section, the term set was constructed respectively and each section had two groups of terms sets based on non-VTE patients and VTE patients separately. Terms which did not exist in candidate words were filtered and four features were calculated for remain terms: Word frequency, Document frequency, Positive and negative times, Entropy of first order neighbors. 5) For Having importance ranking of terms and sections, we proposed an approach to select the section automatically and then build RF model based on picked terms. Just like the greedy feature selection method, every time terms from one section with relatively high AUC (Only) were added to construct new features and RF models were re-trained to try to improve present prediction performance. The selection process stopped when there was no improvement on AUC scores.

Result and Discussions

A. Acknowledgement It is my privilege to acknowledge with deep sense of gratitude to my Project Guide Prof.Dr.Mangesh Ghonge for his valuable suggestions and guidance throughout my course of study and timely help given to me in completion of Dissertation. I gladly take this opportunity to thank Prof.Dr.Amol Potgantwar, Head of Department, for valuable guidance of Dissertation. I would also like to thank Dr.S.T,Gandhe, Principal, for providing facilities during Dissertation work. I am thankful to all those who helped us directly or indirectly for Dissertation work..

B. Figures and Tables

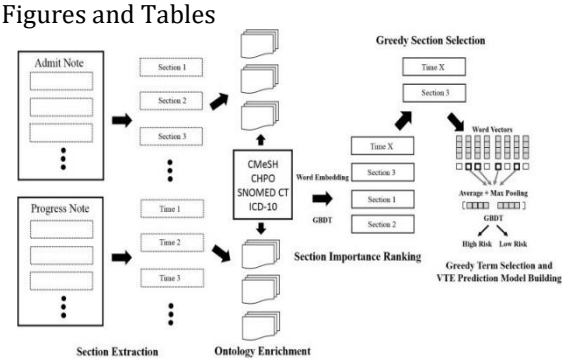

Figure: Architecture Diagram

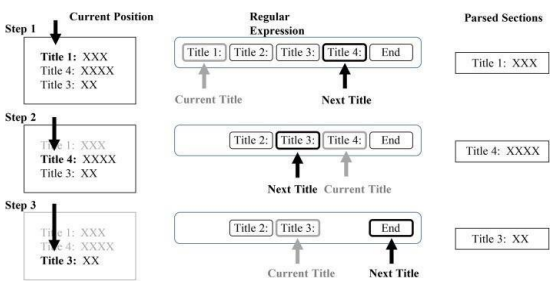

Figure Example of the Greedy Algorithm

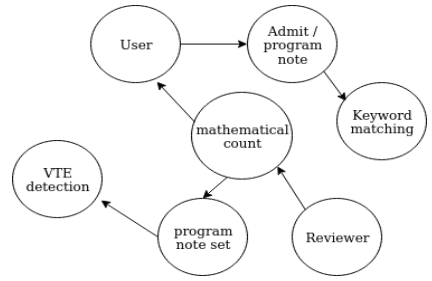

Figure: Use case

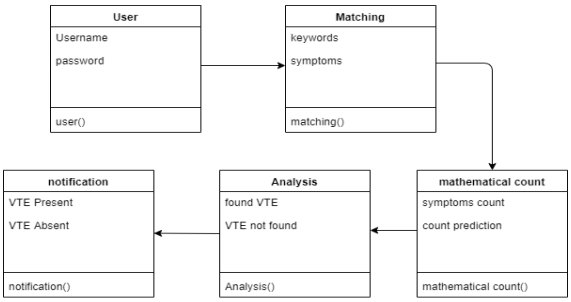

Figure: Class Diagram

Conclusions

In this study, a method of ontology-based VTE risk factors mining and model establishment from medical records is developed and its efficiency is demonstrated on real clinical dataset. Selected terms and sections from medical records help the clinicians discover potential VTE risk factors and RF model built based on these terms improves the performance of VTE prediction. This method is expected to be applied in more diseases and embedded into the EHR system to assist clinical work

References

[1]. Weng, S.F., Reps, J., Kai, J., Garibaldi, J.M., and Qureshi, N.: 'Can machine- learning improve cardiovascular risk prediction using routine clinical data?', Plos One, 2017, 12, (4), pp. e0174944

[2]. Choi, E., Bahadori, M.T., Song, L., Stewart, W.F., and Sun, J.: 'GRAM: Graph- based Attention Model for Healthcare Representation Learning', 2016, pp. 787- 795

[3]. Riccardo, M., Li, L., Kidd, B.A., and Dudley, J.T.:'Deep Patient: An Unsuper- vised Representation to Predict the Future of Patients from the Electronic Health Records', Scientific Reports, 2016, 6, pp.26094

[4]. Casanova, R., Hsu, F.C., Sink, K.M., Rapp, S.R., Williamson, J.D., Resnick, S.M., and Espeland,M.A.: 'Alzheimer's Disease Risk Assessment Using Large- Scale Machine Learning Methods', Plos One, 2013, 8, (11), pp. e7794

[5]. Ferroni, P., Zanzotto, F.M., Scarpato, Ko¨ hler, S., Doelken, S.C., Mungall, C.J., Bauer, S.,Firth, H.V., Bailleul- Forestier, I., Black, G.C.,Brown, D.L., Brudno, M., and Campbell, J.: 'The Hu- man Phenotype Ontology project: linking molecular biology and disease through phenotype data', Nucleic acids research, 2013, 42, (D1), pp.D966-D974

[6]. Barbar, S., Noventa, F., Rossetto, V., Ferrari, A.,Brandolin, B., Perlati, M., De, B.E.,Tormene, D.,Pagnan, A., and Prandoni, P.: 'A risk assessment model for the identification of hospitalized medical patients at risk for venous thromboem- bolism: the Padua Prediction Score', Journal of Thrombosis & Haemostasis Jth, 2010, 8, (11), pp. 2450