© 2017 IJSRST | Volume 3 | Issue 3 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

Drug safety and automation intelligence

Mahima Gupta, Disha Jain, Vinita Gangurde, Jewel Damaralu
Department of Computer Engineering, KKWIEER, Nashik, Maharashtra, India

ABSTRACT

In the area of digitalization, the documentation on patients of health systems (pharmacovigilance) is also being stored in electronic format. Because of this fact the volume of digital information generated in the hospitals is growing exponentially. Professionals often have to manage an excess of data and different kinds of information. The manner in which this sensitive information is presented to the doctors can help in the decision-making process and also alleviate the workload of several services within a hospital. All these facts make the creation of a robust system an important challenge for the Natural Language Processing, Text Mining & Artificial Intelligence research community. In this context the goal of this work is to obtain the Adverse Drug Reactions (ADRs) that are stated in the Electronic Health Records (EHRs) in a robust way. This need arises when experts have to prescribe a drug, since before that, they have to know if the patient has suffered from adverse reactions to substances or drugs. The final system should present the ADRs in the given EHR, showing the drug-disease pairs that triggered each ADR event.
Today, the web has influenced people like never before. If a person wants to search an information, neither does he/she go to the library nor is he/she asks his/her friends and family, as there are many information sites on the web which provide a variety of information. The information is most of the times in either structured, semi structured or unstructured format. Efficient strategies for identification and extraction of information about adverse drug effects from free-text resources are needed to support pharmacovigilance research. Therefore, this work focuses on the adaptation of a machine learning-based relation extraction system for the identification and extraction of drug-related adverse effects from case reports. It relies on a ontology-driven methodology. Qualitative evaluation of the system show robust results.

Keywords : Drug, adverse events, drug safety, artificial intelligence

I. INTRODUCTION

The rapid growth of electronically available health related information (be it in electronic medical records or social media) plus the advances in Natural Language Processing (NLP) and machine learning algorithms present a unique opportunity to massively mine data for the presence of ADR mentions. Prior work has focused on automatic extraction of ADR mentions from electronic medical records (Aramaki, Miura, & Tonoike, 2010) and from user comments in social media (Nikfarjam & Gonzalez, 2011).

Health-related social networking sites are more popular than ever, and are generally accepted as a viable platform to discuss health-related experiences, including symptoms and treatments for different diseases, as well as their side effects. Because of the costs associated with post-marketing ADRs caused by drugs, and the large volume of user posted information available in social media, there is a strong motivation for systems that can automatically monitor social media sites and generate signals when adverse reactions frequently occur for specific drugs.

To Extract ICSR data points and Identify all Adverse Drug Events, Drugs, Medical Condition, Patient Identification, Seriousness Criteria from free Text Electronic Patient Records and Information.

In this solution, input can be a single a PDF, word or text file of size maximum 5 MB, which contains medical case review. The goal of this work is to obtain the Adverse Drug Reactions (ADRs) that are stated in the Electronic Health Records (EHRs) in a robust way. This need arises when experts have to prescribe a drug, since

before that, they have to know if the patient has suffered from adverse reactions to substances or drugs. The final system should present the ADRs in the given EHR, showing the drug-disease pairs that triggered each ADR event. For example, the system should be capable of extracting ADRs such as "As a result of the steroidal treatment, hyperglycemic decomposition was produced which requires treatment with insulinization" from a given EHR, showed in the figure 1. In this case, the disease "hyperglycemic decomposition" has been caused by the "steroidal treatment". The output will be an E2B XML File

II. LITERATURE SURVEY

To Extract ICSR data points and identify all Adverse Drug Events, Drugs, Medical Condition, Patient Identification, Seriousness Criteria from free Text Electronic Patient Records and Information.

The rapid growth of electronically available health related information (be it in electronic medical records or social media) plus the advances in Natural Language Processing (NLP) and machine learning algorithms present a unique opportunity to massively mine data for the presence of ADR mentions. Prior work has focused on automatic extraction of ADR mentions from electronic medical records (Aramaki, Miura, & Tonoike, 2010) and from user comments in social media (Nikfarjam & Gonzalez, 2011).

Health-related social networking sites are more popular than ever, and are generally accepted as a viable platform to discuss health-related experiences, including symptoms and treatments for different diseases, as well as their side effects. Because of the costs associated with post-marketing ADRs caused by drugs, and the large volume of user posted information available in social media, there is a strong motivation for systems that can automatically monitor social media sites and generate signals when adverse reactions frequently occur for specific drugs.

Most of the previous text mining research related to pharmacovigilance is focused on electronic health records (Aramaki et al., 2010; Friedman, 2009; Wang et al., 2009), and medical case reports (Gurulingappa, Rajput, & Toldo, 2012; Toldo, Bhattacharya, & Gurulingappa, 2012). Harpaz et al. (2012) provide a thorough survey on the existing approaches for post-marketing pharmacovigilance, exploring various resources such as electronic health records, spontaneous adverse drug reporting systems and biomedical literature. Social media was relatively unexplored for this purpose until recently. Leaman et al. (2010) analyzed user comments in social media and demonstrated that the comments contain extractable drug safety information.

The authors used a hybrid lexicon and rule-based system for ADR concept extraction. Nikfarjam & Gonzalez (2011) proposed a 3 http://sideeffects.embl.de/ pattern-based technique based on association rule mining, which extracts ADR mentions based on the language patterns used by patients in social media for expressing ADRs. In a recent study Yates & Goharian (2013) analyzed the value of user comments in revealing the unknown adverse effects by evaluating the extracted ADRs against the SIDER database3 which contains information about the known adverse effects. There are similar studies for automatic ADR mention extraction, targeting online patient discussions (Yates & Goharian, 2013; Benton et al., 2011; Sampathkumar, Luo, & Chen, 2012). While these techniques can be used to extract ADR mentions from the available online user contents, our task only requires a binary decision about the comment being ADR or NoADR. Chee et al. (2011) classified user posts on online groups to predict the candidate FDA watch list drugs for further investigation with regards to drug safety. They used an ensemble based classification technique to identify drugs that are likely to be in the watch list category. Our work is different in two ways; first, our dataset is from health related social network, which generally contains unstructured sentences, incorrect spellings, and more informal language compared to electronic health records. Secondly, we hypothesize that a drug can be classified as watch list (we refer to these as black box) or normal based on the amount of adverse events that are reported about the drug.

III. IMPLEMENTATION

BASICS CONCPTS OF ICSR Data Extraction & Validity Classification

The system uses external English Natural Language Processing tools for pre-processing the search query entered by user. The fundamental use of these tools is to identify the Nouns, Proper Nouns present in the search query. It returns a list of nouns that are present in the query. There two important tools which are used i.e. the Stanford Parser and the Stanford PoS Tagger. The PoS Tagger is a tool which simply marks each word based on type of English Part of Speech. It never looks for the semantics of the word i.e. where exactly the word occurs, what are predecessor and successor words, what could be the meaning, etc. Since, Natural Language Processing has several ways of marking PoS tags i.e. by following tags defined in Penn Tree Bank, British Corpus or any other standard PoS Tag set. There might be a case that a word may be classified as Noun when used with Penn Tree Bank tag set but might be of any other type when classified using other Tagset. So, this might be ambiguous as if the word is not noun and if it is classified as noun then it might make the search process ambiguous and may result in improper extraction. Another crucial tool used is the Parser. It is a deep parser which looks at search query as a proper English sentence. It processes the query in three phases and then based on the semantics or context of the sentence, the parse is drawn and it marks each token accordingly with PoS tag. Since, it looks at the semantics of the sentence and then decided what could be a PoS tag might be a bit ambiguous as most of the times users enter random words in search query as they don't have exact sentence that could help search the desired data. So, the parser could find the sentence syntactically wrong or could build altogether different meaning which could lead to the improper parsing of the search query. This could lead to the improper search by Machine Learning and also could affect the accuracy of the search.

Machine Learning based Classification

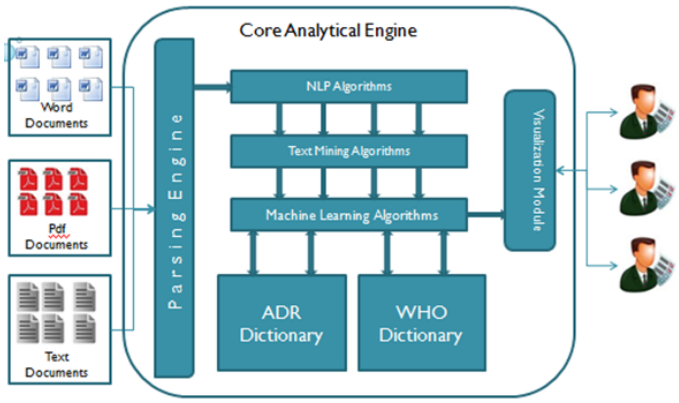
Machine learning – subfield of computer science (more particularly soft computing) that evolved from the study of pattern recognition and computational learning theory in artificial intelligence. In 1959, Arthur Samuel defined machine learning as a "Field of study that gives computers the ability to learn without being explicitly programmed".[2]Machine learning explores the study and construction of algorithms that can learn from and make predictions on data. Such algorithms operate by building a model from an example training set of input observations in order to make data-driven predictions or decisions expressed as outputs, rather than following strictly static program instructions.

In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations)new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. An example would be assigning a given email into "spam" or "non-spam" classes or assigning a diagnosis to a given patient as described by observed characteristics of the patient (gender, blood pressure, presence or absence of certain symptoms, etc.). Classification is an example of pattern recognition.
In the terminology of machine learning, classification is considered an instance of supervised learning, i.e. learning where a training set of correctly identified observations is available. The corresponding unsupervised procedure is known as clustering, and involves grouping data into categories based on some measure of inherent similarity or distance.

IV. PROPOSED SYSTEM

In this system various types of files such as PDF, word and txt can be uploaded as a input file. The Input file will passed through parsing engine which will extract the raw text from the input file. This Raw Text is used as a input for NLP Algorithm in which different NLP tasks like Tokenization, Steaming, Stop-Word Removal etc. are performed that will help in enriching the raw text. The pre-proceed raw text will be used to mine the ICSR Data points from CASE Reviews using WHO Drug and ADR dictionaries .

Then the machine learning algorithm is used to classify documents as valid and invalid case reviews. The Mined ICSR data points will be incorporated in the standard E2B xml and will be displayed to use through visualization layer and the user will also be able to download the E2B xml file.



The solution inputs single PDF, word or text file of size maximum 5 MB , which contains medical case review. the goal of this work is to obtain the Adverse Drug Reactions (ADRs), that are stated in the Electronic Health Records (EHRs) in a robust way. This need arises when experts have to prescribe a drug, since before that, they have to know if the patient has suffered from adverse reactions to substances or drugs. The final system should present the ADRs in the given EHR, showing the drug-disease pairs that triggered each ADR event. For example, the system should be capable of extracting ADRs such as "As a result of the steroidal treatment, hyperglycemic decompensation was produced which requires treatment with insulinization" from a given EHR, showed in the figure 1. In this case, the disease "hyperglycemic decompensation" has been caused by the "steroidal treatment". The output will be an E2B XML File.

V. CONCLUSION

The solution inputs single PDF, word or text file of size maximum 5 MB , which contains medical case review. the goal of this work is to obtain the Adverse Drug Reactions (ADRs), that are stated in the Electronic Health Records (EHRs) in a robust way. This need arises when experts have to prescribe a drug, since before that, they have to know if the patient has suffered from adverse reactions to substances or drugs. The final system should present the ADRs in the given EHR, showing the drug-disease pairs that triggered each ADR event. For example, the system should be capable of extracting ADRs such as "As a result of the steroidal treatment, hyperglycemic decompensation was produced which requires treatment with insulinization" from a given EHR, showed in the figure 1. In this case, the disease "hyperglycemic decompensation" has been caused by the "steroidal treatment". The output will be an E2B XML File.

VI. REFERENCES

[1]. Feng, J. And Li, G., "Efficient Fuzzy Type-Ahead Search in XML Data," Proc. of IEEE Transactions on Knowledge And Data Engineering, Vol. 24 No. 5, pp. 882-895, 2012.

[2]. Aramaki, E., Miura, Y., Tonoike, M., Ohkuma, T., Masuichi, H., Waki, K., and Ohe, K.(2010). Extractionof adverse drug effects from clinical records. In Studies HealthTechnology Informatics, volume 160, pages 739–743.

[3]. Benton, A., Ungar, L., Hill, S., Hennessy, S., Mao, J., Chung, A., Leonard, C., and Hol-mes, J. (2011). Identifying potential adverse effects using the web:A new approach to medical hypothesis generation. Journal of Biomedical Informatics, 44, 989–996.

[4]. Burges, C. (1998). A tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 2.

[5]. Delamarre, D., Lillo-Le Louet, A., Jamte, A., Sadou, E., Ouazine, T., Burgun, A., and Jaulent, M. (2010). Documentation in pharmacovigilance: using an ontology to extend and normalize Pubmed queries. In Studies Health Technology Informatics, volume 160, pages 518–522.

[6]. Giuliano, C., Lavelli, A., Pighin, D., and Romano, L. (2007). FBK-IRST: Kernel Methods for Semantic Relation Extraction. In Proceedings of the Fourth International Workshop on Semantic Evaluations.

[7]. Gurulingappa, H., Fluck, J., Hofmann-Apitius, M., and Toldo, L. (2011). Identification of adverse drug event assertive sentences in medical case reports. InFirst International Workshop on Knowledge Discovery and Health Care Management (KD-HCM), European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD).

[8]. Gurulingappa, H., Mateen-Rajput, A., Roberts, A., Fluck, J., Hofmann-Apitius, M., and Toldo, L.

(2012). Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. Journal of Biomedical Informatics.

[9]. Hanisch, D., Fundel, K., Mevissen, H.-T., Zimmer, R., and Fluck, J. (2005). Prominer: rule-based protein and gene entity recognition. BMC Bioinformatics, 6 Suppl 1, S14.

[10]. Hauben, M. and Bate, A. (2009). Decision support methods for the detection of adverse events in post-marketing data. Drug Discov Today, 14(7-8), 343–357. Henegar, C., Bousquet, C., Lillo-Le Louet, A., Degoulet, P., and Jaulent, M.-C. (2006).

[11]. Building an ontology of adverse drug reactions for automated signal generation in pharmacovigilance. Computers in Biology and Medicine, 36, 748–767.Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A. C., and Wishart, D. S. (2011).

[12]. Drugbank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res, 39(Database issue), D1035–D1041. Leaman, R., Wojtulewicz, L., Sullivan, R., Skariah, A., Yang, J., and Gonzalez, G. (2010).

[13]. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. In Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, pages 117–125. Merrill, G. H. (2008). The meddra paradox. AMIA Annu Symp Proc, pages 470–474.