International Journal of Advanced Information Science and Technology (IJAIST) ISSN: 2319:2682
Vol.3, No.8, August 2014 DOI:10.15693/ijaist/2014.v3i8.84-89

Mining Electronic health Records (EHR) Using Machine Learning Approach

J Gnana Jeslin[1]

[1]*Assistant Professor, Jeppiaar Engineering College*

***Abstract*— The Machine Learning (ML) field has gained its momentum in almost any domain of research and just recently has become a reliable tool in the medical domain. The empirical domain of automatic learning is used in tasks such as medical decision support, medical imaging, protein-protein interaction, extraction of medical knowledge, and for overall patient management care. ML is envisioned as a tool by which computer-based systems can be integrated in the healthcare field in order to get a better, more efficient medical care. This paper describes a ML-based methodology for building an application that is capable of identifying and disseminating healthcare information. This paper is focused mainly on two tasks: automatically identifying sentences published in medical abstracts - Medline as containing or not information about diseases and treatments, and automatically identifying semantic relations that exist between diseases and treatments, as expressed in these texts. The second task is focused on three semantic Relations: Cure, Prevent, and Side Effect.**

***Keywords*— Electronic health records, health care, machine learning, natural language processing.**

I. Introduction

People care deeply about their health and want to be, now more than ever, in charge of their health and health care. Life is more hectic than has ever been; the medicine that is practiced today is an Evidence-Based Medicine in which medical expertise is not only based on years of practice but on the latest discoveries as well. Tools that can help us manage and better keep track of our health such as Google Health and Microsoft Health Vault are reasons and facts that make people more powerful when it comes to health care knowledge and management. The traditional health care system is also becoming one that embraces the Internet and the electronic world.

Electronic Health Record is becoming the standard in the healthcare domain. Researches and studies show that the potential benefits of having an EHR system are: Health information recording and clinical data repositories—immediate access to patient diagnoses, allergies, and lab test results that enable better and time-efficient Medical decisions; Medication management—rapid access to information regarding potential adverse drug reactions, immunizations, supplies, etc.

Decision support—the ability to capture and use quality medical data for decisions in the workflow of healthcare; and Obtain treatments that are tailored to specific health needs—rapid access to information that is focused on certain topics. In order to embrace the views that the EHR system has, we need better, faster, and more reliable access to information. In the medical domain, the richest and most used source of information is Medline, a database of extensive Life science published articles. All research discoveries come and enter the repository at high rate, making the process of identifying and disseminating reliable information a very difficult task.

The work that we present in this paper is focused on two tasks: automatically Identifying sentences published in medical abstracts (Medline) as containing or not information about diseases and treatments, and automatically identifying semantic relations that exist between diseases and treatments, as expressed in these texts. The second task is focused on three semantic relations: Cure, Prevent, and Side Effect

II. Related Work

Entity recognition for Diseases and Treatments-- The most relevant work is done by Rosario and Hearst[2].It uses Hidden Markov Models and maximum entropy models to perform both the task of entity recognition and the relation discrimination. Their representation techniques are based on words in context, part of speech information, phrases, and a medical lexical ontology—Mesh terms.The task of relation extraction or relation identification is previously done by (Craven, [3])with a focus on biomedical task, gene disorder association(Ray and Craven, [4]) and diseases and drugs(Shrinivasan and Rindflesch, [5]). Usually, the data sets used in biomedical specific tasks use short texts, often sentences. This is the case of the first two related works mentioned above. The tasks often entail identification of relations between entities that co-occur in the same sentence.

Rule-based approaches

It has been widely used for solving relation extraction tasks. The main sources of information used by this technique are either syntactic: part-of speech (POS) and syntactic structures; or semantic information in the form of fixed patterns that contain words that trigger a certain relation. The best rule-based systems are the ones that use rules constructed manually or semi automatically—extracted automatically and refined manually. A positive aspect of rule-based systems is the fact that they obtain good precision results, while the recall levels tend to be low. They tend to require more human-expert effort than data-driven methods (though human effort is needed in data-driven methods too, to label the data).

The semantic rule-based approaches suffer from the fact that the lexicon changes from domain to domain, and new rules need to be created each time. Certain rules are created for biological corpora, medical corpora, pharmaceutical corpora, etc. Systems based on semantic rules applied to full-text articles are described by Friedman et al. [6], on sentences by Pustejovsky et al. [7], and on abstracts by Rindflesch et al. [5]. Some researchers combined syntactic and semantic rules from Medline abstracts in order to obtain better systems with the flexibility of the syntactic information and the good precision of the semantic rules, e.g., Gaizauskas et al. [8]

III. The Proposed Approach

A. Tasks and Data Sets

The two tasks that are undertaken in this paper provide the basis for the design of an information technology framework that is capable to identify and disseminate healthcare information. The first task identifies and extracts informative sentences on diseases and treatments topics, while the second one performs a finer grained classification of these sentences according to the semantic relations that exists between diseases and treatments.

The first task (task 1 or sentence selection) identifies sentences from Medline published abstracts that talk about diseases and treatments. The task is similar to a scan of sentences contained in the abstract of an article in order to present to the user-only sentences that are identified as containing relevant information (disease treatment information).

The second task (task 2 or relation identification) has a deeper semantic dimension and it is focused on identifying disease-treatment relations in the sentences already selected as being informative (e.g., task 1 is applied first). We focus on three relations: Cure, Prevent, and Side Effect, a subset of the eight relations that the corpus is annotated with. We decided to focus on these three relations because these are most represented in the corpus while for the other five, very few examples are available. Table 1 presents the original data set, the one used by Rosario and Hearst [2], that we also use in our research. The numbers in parentheses represent the training and test set size. For example, for Cure relation, out of 810 sentences present in the data set, 648 are used for training and 162 for testing. The approach used to solve the two proposed tasks is based on NLP and ML techniques. In a standard supervised ML setting, a training set and a test set are required. The training set is used to train the ML algorithm and the test set to test its performance.

The task of identifying the three semantic relations is addressed in two ways:

Setting 1. Three models are built. Each model is focused on one relation and can distinguish sentences that contain the relation from sentences that do not. This setting is similar to a two-class classification task in which instances are labeled either with the relation in question (Positive label) or with nonrelevant information (Negative label);

Setting 2. One model is built, to distinguish the three relations in a three-class classification task where each sentence is labeled with one of the semantic relations

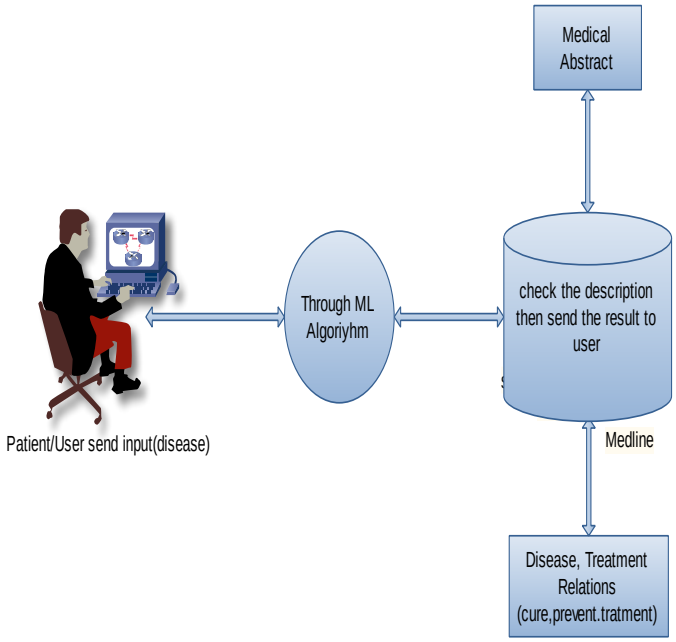

Figure 2: Existing System Architecture

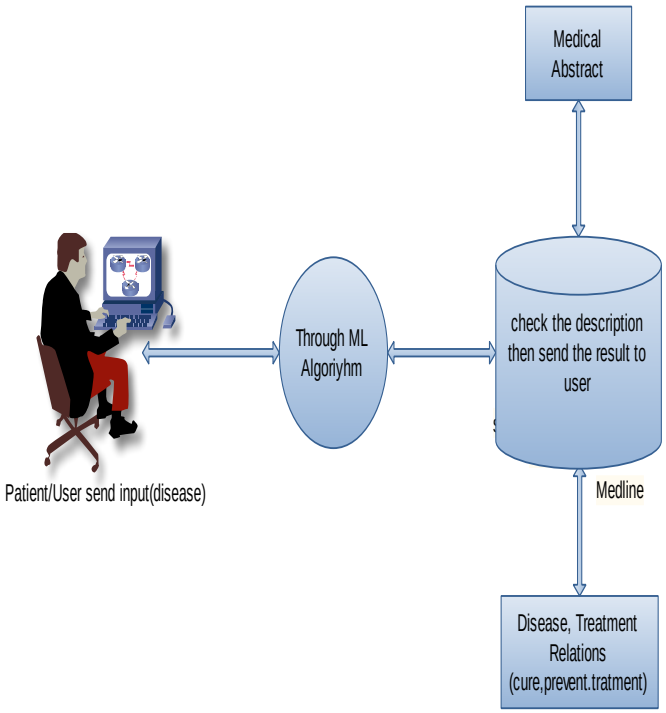

Figure1: Proposed System Architecture

TABLE 1
Data Set Description, Taken from Rosario and Hearst ('04)

Relationship	Definition and Example
Cure 810 (648, 162)	TREAT cures DIS *Intravenous immune globulin for recurrent spontaneous abortion*
Only DIS 616 (492, 124)	TREAT not mentioned *Social ties and susceptibility to the common cold*
Only TREAT 166 (132, 34)	DIS not mentioned *Flucticasome propionate is safe in recommended doses*
Prevent 63 (50, 13)	TREAT prevents the DIS *Statins for prevention of stroke*
Vague 36 (28, 8)	Very unclear relationship *Phenylbutazone and leukemia*
Side Effect 29 (24, 5)	DIS is a result of a TREAT *Malignant mesodermal mixed tumor of the uterus following irradiation*
NO Cure 4 (3, 1)	TREAT does not cure DIS *Evidence for double resistance to permethrin and malathion in head lice*
Total relevant: 1724 (1377, 347)	
Irrelevant 1771 (1416, 355)	Treat and DIS not present *Patients were followed up for 6 months*
Total: 3495 (2793, 702)	

In brackets, are the numbers of instances used for training and for testing, respectively

B. Classification Algorithms and Data Representations:-

In ML, as a field of empirical studies, the acquired expertise and knowledge from previous research guide the way of solving new tasks. The models should be reliable at identifying informative sentences and discriminating disease- treatment semantic relations. The research experiments need to be guided such that high performance is obtained. The experimental settings are directed such that they are adapted to the domain of study (medical knowledge) and to the type of data we deal with (short texts or sentences), allowing for the methods to bring improved performance. There are at least two challenges that can be encountered while working with ML techniques. One is to find the most suitable model for prediction. The ML field offers a suite of predictive models (algorithms) that can be used and deployed. The task of finding the suitable one relies heavily on empirical studies and knowledge expertise. The second one is to find a good data representation and to do feature engineering because features strongly influence the performance of the models. Identifying the right and sufficient features to represent the data for the predictive models, especially when the source of information is not large, as it is the case of sentences, is a crucial aspect that needs to be taken into consideration. These challenges are addressed by trying various predictive algorithms, and by using various textual representation techniques that we consider suitable for the task. As

classification algorithms, we use a set of six representative models: decision-based models (Decision trees), probabilistic models (Naı¨ve Bayes (NB) and Complement Naı¨ve Bayes (CNB), which is adapted for text with imbalanced class distribution), adaptive learning (Ada-Boost), a linear classifier (support vector machine (SVM) with polynomial kernel), and a classifier that always predicts the majority class in the training data (used as a baseline). We decided to use these classifiers because they are representative for the learning algorithms in the literature and were shown to work well on both short and long texts. Decision trees are decision-based models similar to the rule-based models that are used in handcrafted systems, and are suitable for short texts. Probabilistic models, especially the ones based on the Naı¨ve Bayes theory, are the state of the art in text classification and in almost any automatic text classification task. Adaptive learning algorithms are the ones that focus on hard-to-learn concepts, usually underrepresented in the data, a characteristic that appears in our short texts and imbalanced data sets. SVM-based models are acknowledged state-of-the-art classification techniques on text. All classifiers are part of a tool called Weka.[9] (Oana Frunza[1] et al).One can imagine the steps of processing the data (in our case textual information—sentences) for ML algorithms as the steps required to obtain a database table that contains as many columns as the number of features selected to represent the data, and as many rows as the number of data points from the collection (sentences in our case).

(i). Bag-of-Words Representation:-

The bag-of-words (BOW) representation is commonly used for text classification tasks. It is a representation in which features are chosen among the words that are present in the training data. Selection techniques are used in order to identify the most suitable words as features. After the feature space is identified, each training and test instance is mapped to this feature representation by giving values to each feature for a certain instance. Two most common feature value representations for BOW representation are: binary feature values—the value of a feature can be either 0 or 1, where 1 represents the fact that the feature is present in the instance and 0 otherwise; or frequency feature values—the value of the feature is the number of times it appears in an instance, or 0 if it did not appear.

(ii). NLP and Biomedical Concepts Representation:-

The second type of representation is based on syntactic information: noun-phrases, verb-phrases, and biomedical concepts identified in the sentences. In order to extract this type of information, we used the Genia11 tagger tool. The tagger analyzes English sentences and outputs the base forms, part-of-speech tags, chunk tags, and named entity tags. The tagger is specifically tuned for biomedical text such as Medline abstracts. The noun and verb-phrases identified by the tagger are features used for the second representation technique.

*(iii). Medical Concepts (UMLS) Representation***:-**

In order to work with a representation that provides features that are more general than the words in the abstracts (used in the BOW representation), we also used the Unified Medical Language system concept representations. UMLS is a knowledge source developed at the US National Library of Medicine) and it contains a met thesaurus, a semantic network, and the specialist lexicon for biomedical domain. The met thesaurus is organized around concepts and meanings; it links alternative names and views of the same concept and identifies useful relationships between different concepts.

C. EHR System:-

An electronic health record (EHR) is an evolving concept defined as a systematic collection of electronic health information about individual patients or populations. It is a record in digital format that is theoretically capable of being shared across different health care settings. In some cases this sharing can occur by way of network-connected enterprise-wide information systems and other information networks or exchanges. EHRs may include a range of data, including demographics, medical history, medication and allergies, immunization status, laboratory test results, radiology images, vital signs, personal stats like age and weight, and billing information.

IV. Validating The Proposed System

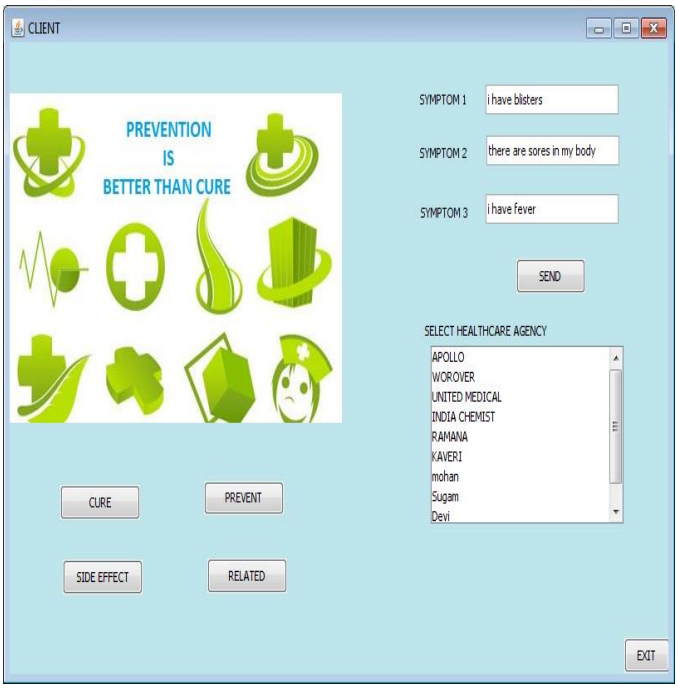

Figure3: Clinical Decision Support System

Figure 3 shows a clinical decision made where the user gives a symptom and the system displays the most related disease based on the symptom keyword

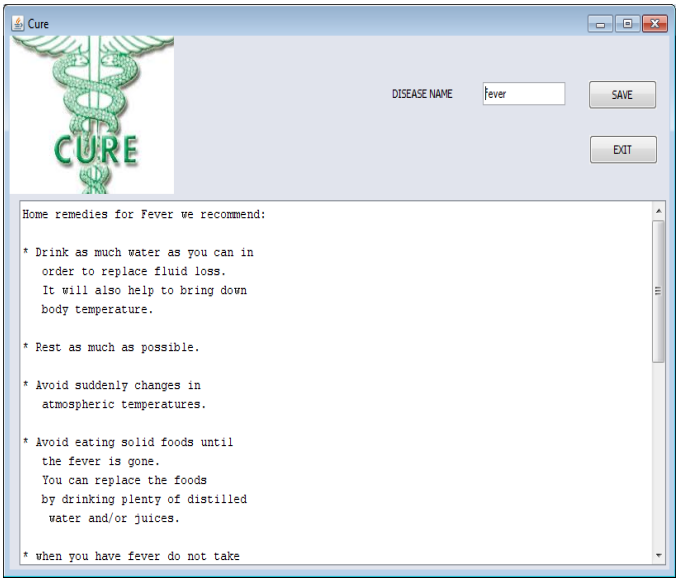

Figure 4: Cure For The Disease Of The Client

Figure 4 shows the cure for the disease based on the symptoms given by the client.

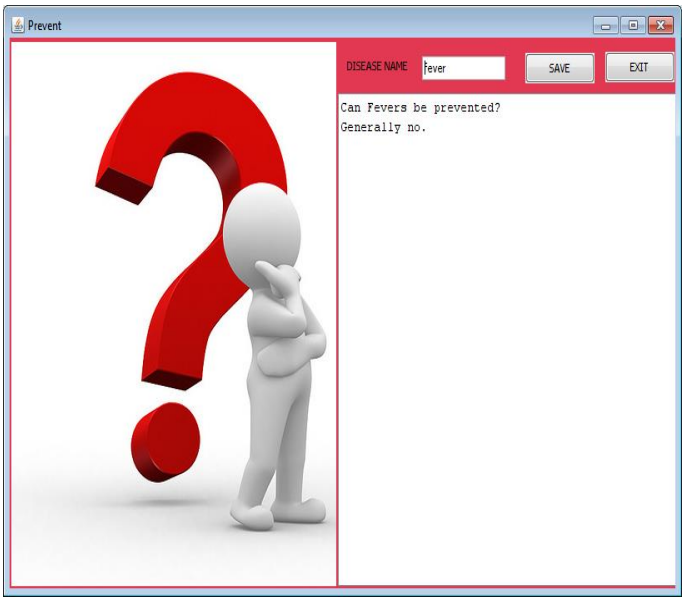

Figure 5: Prevention For The Disease Of The Client

Figure 5 shows the prevention for the disease based on the symptoms given by the client.

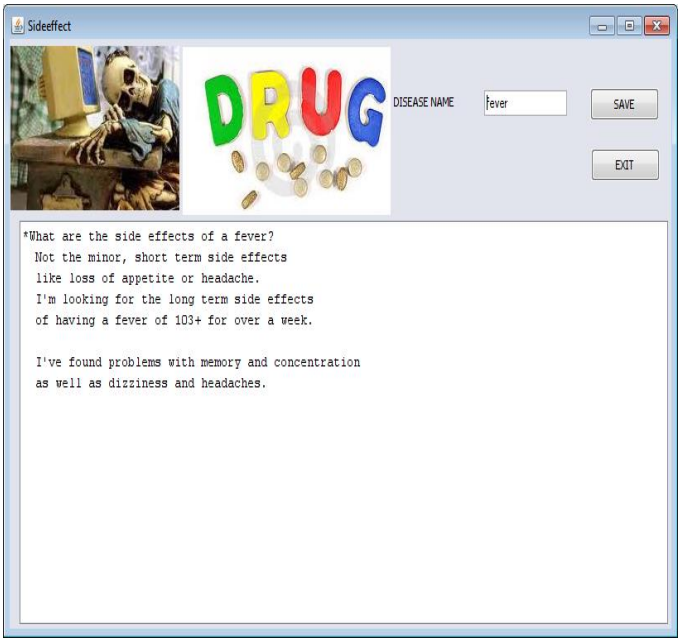

Figure 6: Side Effect For The Disease Of The Client

Figure 6 shows the various side effects for the disease based of the client.

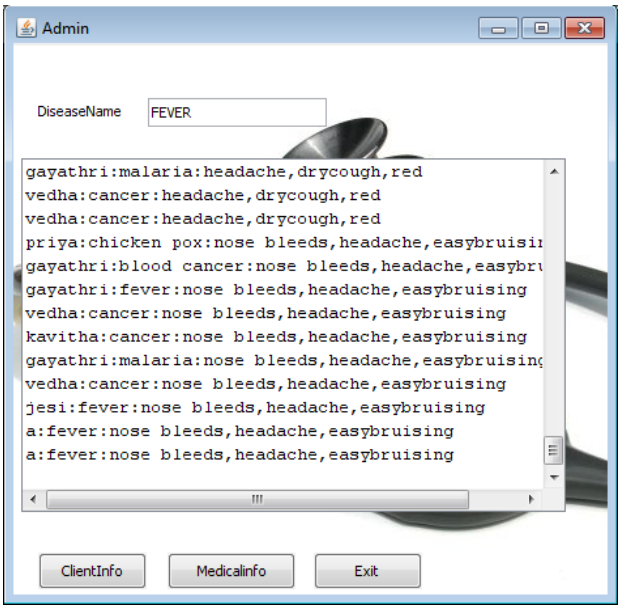

Figure 7: Maintaining Electronic Patient Record

Figure 7 gives the details about the records collected in the EPR along with patient name symptoms etc.

V. Conclusion

In this paper we are mining the data from EHR and medical abstracts. From the results arrived from mining the text database we are diagnosing the disease of the patients, prescribing them medicines, giving details about the side effects of the disease. This system also maintains the database of the various symptoms given by the users and

can find the maximum occurred symptom of the patients. This also helps to correlate between the various different symptoms.

VI. References

[1] Oana Frunza, Diana Inkpen, and Thomas Tran " A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts" vol. 23, 2011.

[2] B. Rosario and M.A. Hearst, "Semantic Relations in Bioscience Text," Proc. 42nd Ann. Meeting on Assoc. for Computational Linguistics, vol. 430, 2004.

[3] M. Craven, "Learning to Extract Relations from Medline," Proc. Assoc. for the Advancement of Artificial Intelligence, 1999.

[4] S. Ray and M. Craven, "Representing Sentence Structure in Hidden Markov Models for Information Extraction," Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI '01), 2001.

[5] P. Srinivasan and T. Rindflesch, "Exploring Text Mining from Medline," Proc. Am. Medical Informatics Assoc. (AMIA) Symp., 2002.

[6] C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky, "GENIES: A Natural Language Processing System for the Extraction of Molecular Pathways from Journal Articles," Bioinformatics, vol. 17, pp.S74-S82, 2001.

[7] J. Pustejovsky, J. Castan˜ o, J. Zhang, M. Kotecki, and B. Cochran, "Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relations," Proc. Pacific Symp. Biocomputing, vol. 7, pp. 362- 373, 2002.

[8] R. Gaizauskas, G. Demetriou, P.J. Artymiuk, and P. Willett, "Protein Structures and Information Extraction from Biological Texts: The PASTA System," Bioinformatics, vol. 19, no. 1, pp. 135- 143, 2003.

[9] Weka tool,

http://www.cs.waikato.ac.nz/ml/weka/.

[10]Microsoft Health Vault,

http://healthvault.com

[11] Health care tracker,

http://healthcaretracker.wordpress.com/

[12] Medline Database,

http://www.proquest.com/en US/catalogs/databases/detail/medline_ft.s html

[13] Medical Subject Headings, http://www.nlm.nih.gov/mesh/meshhome.html.

[14] Google health report,

https://www.google.com/health