ISSN 2250-3153

Healthcare Information Using Machine Learning Approach

R. Buvaneshwari, K. Lavanya, R. Vanitha

Department of Information Technology,
Periyar maniammai University, Vallam, Thanjavur, India

Abstract- In recent year we develop Machine Learning (ML) approach.ML is to build computer systems that can adapt and learn from their experience.ML is the domain of research and recently it has develop in medical domain .The domain is automatically learn some task of healthcare information, medical management, patient health management etc.,. Healthcare deals with the resource, devices and method require optimizing storage, retrieval and use of information in health and bio medicine. Healthcare diagnosis, treatment and prevention of disease, illness, injury in human. This paper ML methodology capable of identifying, spreading the information in healthcare. The extract sentence published from medical paper that mention on disease treatment. The result obtains reliable outcome integrated in medical domain.

Index Terms - Healthcare, 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 healthcare. No clinician would consider entering clinical practice without knowing the rudiments of history-taking and physical examination, nor would clinicians consider independent practice without a basic understanding of how the drugs they prescribe act on their patients. Yet, traditionally, clinicians have started practice without an ability to understand evidence about how they should interpret what they find on history and physical examination, or the magnitude of the effects they might expect when they offer patients medication.

Evidence-based medicine (EBM) is aims to apply the best available evidence gained from the scientific method to clinical decision making.[1] It seeks to assess the strength of evidence of the risks and benefits of treatments (including lack of treatment) and diagnostic tests. It is the integration of best research evidence with clinical expertise and patient values. It aims to apply the best available evidence gained from the scientific method to medical decision making and it seeks to assess the quality of evidence of the risks and benefits of treatments. Tools that can help us manage and better keep track of our health such as Google Health[1] was a personal health information centralization service (sometimes known as personal health record services) by Google and Microsoft HealthVault[2] is a web-based platform from Microsoft to store and maintain health and fitness information are reasons and facts that make people more powerful when it comes to healthcare knowledge and management.

The traditional healthcare system is also becoming one that embraces the Internet and the electronic world. An Electronic Health Record (EHR) (also electronic patient record (EPR) or computerized patient record) is an evolving concept defined as a systematic collection of electronic health information about individual patients or populations.EHR systems are,

A. Health information and data.

Having immediate access to key information - such as patients' diagnoses, allergies, lab test results, and medications - would improve caregivers' ability to make sound clinical decisions in a timely manner.

B. Result management.

The ability for all providers participating in the care of a patient in multiple settings to quickly access new and past test results would increase patient safety and the effectiveness of care.

C. Decision support.

Using reminders prompts, and alerts, computerized decision-support systems would help improve compliance with best clinical practices, ensure regular screenings and other preventive practices, identify possible drug interactions, and facilitate diagnoses and treatments.

D. Patient support.

Tools that give patients access to their health records, provide interactive patient education, and help them carry out home-monitoring and self-testing can improve control of chronic conditions, such as diabetes.

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[4] enables anyone to query the NLM computer's store of journal article references on specific topics. It currently contains 9 million references going back to the mid-1960s a database of extensive life science published articles. All research discoveries come and enter the repository at high rate (Hunter and Cohen[12]), 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.*

The tasks that are addressed here are the foundation of an information technology framework that identifies and spread the healthcare information. People want fast access to reliable

information and in a manner that is suitable to their habits and workflow. Medical care related information (e.g., published articles, clinical trials, news, etc.) is a source of power for both healthcare providers and laypeople. Studies reveal that people are searching the web and read medical related information in order to be informed about their health. Ginsberg et al. [10] show how a new outbreak of the influenza virus can be detected from search engine query data.

Our objective for this work is to show what Machine Learning (ML) techniques—what representation of information and what classification algorithms—are suitable to use for identifying and classifying relevant medical information in short texts. Another objective of is Natural Language Processing (NLP) techniques — is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages.

We envision the potential and value of the findings of our work as guidelines for the performance of a framework that is capable to find relevant information about diseases and treatments in a medical domain repository. The results that we obtained show that it is a realistic scenario to use ML techniques to build a tool, similar to an RSS feed, capable to identify and disseminate textual information related to diseases and treatments. Therefore, this study is aimed at designing and examining various representation techniques in combination with various learning methods to identify and extract biomedical relations from literature.

The contributions that we bring with our work stand in the fact that we present an extensive study of various ML algorithms and textual representations for classifying short medical texts and identifying semantic relations between two medical entities: diseases and treatments. From an ML point of view, we show that in short texts when identifying semantic relations between diseases and treatments a substantial improvement in results is obtained when using a hierarchical way of approaching the task (a pipeline of two tasks).

II. RELATED WORK

The most relevant related work is the work done by Rosario and Hearst [25]. The authors of this paper are the ones who created and distributed the data set used in our research. The data set consists of sentences from Medline [5] abstracts annotated with disease and treatment entities and with eight semantic relations between diseases and treatments. For examine the problem of distinguishing among seven relation types that can occur between the entities “treatment” and “Disease” in bioscience text, and the problem of identifying such entities. The main focus of their work is on entity recognition for diseases and treatments. The authors use Hidden Markov Models and maximum entropy models to perform both the task of entity recognition and the relation discrimination.

The tasks addressed in our research are information extraction and relation extraction. From the wealth of research in these domains, we are going to mention some representative works. The task of relation extraction or relation identification is previously tackled in the medical literature, but with a focus on biomedical tasks: sub cellular-location (Craven [4]), propose an approach to representing the grammatical structure of sentences in the states of the model. Gene-disorder association (Ray and Craven [23]), and diseases and drugs (Srinivasan and Rindflesch, [26]). 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.

There are three major approaches used in extracting relations between entities: co-occurrences analysis, rule-based approaches and statistical methods. The co-occurrences methods are mostly based only on lexical knowledge and words in context, and even though they tend to obtain good levels of recall, their precision is low. Good representative examples of work on Medline abstracts include Jenssen et al. [14] and Stapley and Benoit [27]The extracted information may be referential as for example the names of cellular locations or the names of drugs.

In biomedical literature, rule-based approaches have 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. One of the drawbacks of using methods based on rules is that 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).

Syntactic rule-based relation extraction systems are Complex systems based on additional tools used to assign POS tags or to extract syntactic parse trees. Representative works on syntactic rule-based approaches for relation extraction in Medline abstracts and full-text articles are presented by Thomas et al. [28], Yakushiji et al. [29], and Leroy et al. [16].

Even though the syntactic information is the result of tools that are not 100 percent accurate, success stories with these types of systems have been encountered in the biomedical domain.

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] and Novichkova et al. [20].

Rules are automatically extracted by the learning algorithm when using statistical approaches to solve various tasks. In general, statistical techniques can perform well even with little training data. Various learning algorithms have been used for the statistical learning approach with kernel methods being the popular ones applied to Medline abstracts (Li et al. [17]). In the later mentioned domains, Goadrichet al. [11] IE is the process of finding facts in unstructured text, such as biomedical journals, and putting those facts in an organized system., while Ould et al.[21]experimented with bag-of-word features on sentences. Our work differs from the ones mentioned in this section by the fact that we combine different textual representation techniques for various ML algorithms.

The importance of extracting biomedical information from scientific publications is well recognized. A number of information extraction systems for the biomedical domain have been reported, but none of them have become widely used in practical applications. Most proposals to date make rather simplistic assumptions about the syntactic aspect of natural language.

III. SYSTEM ARCHITECTURE

The system build upon a scalable system developed with the goal of detecting harmful URLs on the web. In this section, we describe our extensions to this system to collect and analyze malicious network activity occurring after infection. To provide context for our research, It first give a brief overview of the overall system described in depth in prior work [10]. This is followed by a detailed description of the light-weight responders which allow us to automatically capture the network-level activity of drive-by downloads.

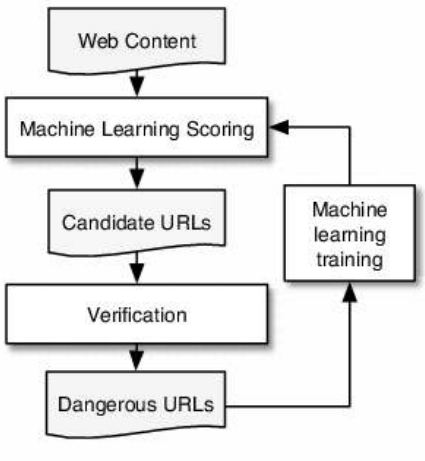

Fig1: Overall system architecture.

Using machine learning techniques, suspicious URLs are selected among billions of web pages for verification in a virtual machine. Our system consists of an efficient first-pass filter followed by a verification component. Figure 1 provides an overview of the system components and their interaction. The first-pass filter is essentially a *map reduce* [5] over billions of web documents. For each web page, we extract several features, including links to known malware ``distribution" sites, suspicious HTML elements, or the presence of code obfuscation. The combination of these features is scored by a model trained on a specialized machine-learning system [3]. URLs with a high score are considered potentially malicious, and are submitted to the verification component for further analysis. The URLs that are verified to be malicious are then exported to Google Web Search and, via the Google Safe Browsing API [1], to other clients. The verification results are also used to retrain the machine learning model. Web content is the textual, visual or aural content that is encountered as part of the user experience on websites. It may include, among other things: text, images, sounds, videos and animations. Even though we may embed various protocols within web pages, the "web page" composed of "html". Content is still the dominant way whereby we share content. In this section we generalize the methodology to *machine learning* of the scoring function. We considered a case where we had to combine Boolean indicators of relevance; here we consider more general factors to further develop the notion of *machine-learned relevance*. Then it goes to Uniform Resource Locator or Universal Resource Locator (URL) candidate a specific character string that constitutes a reference to an Internet resource. A URL is technically a type of Uniform Resource Identifier (URI) but in many technical documents and verbal discussions URL is often used as a synonym for URI. To verify the candidate if it is a dangerous URL it should be trained at machine learning otherwise URL is verified.

IV. PROPOSED SYSTEM

In this paper it have 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 problems addressed in this paper form the building blocks of a framework that can be used by healthcare providers (e.g., private clinics, hospitals, medical doctors, etc.), companies that build systematic reviews (hereafter, SR), or lay people who want to be in charge of their health by reading the latest life science published articles related to their interests. The final product can be envisioned as a browser plug-in or a desktop application that will automatically find and extract the latest medical discoveries related to disease-treatment relations and present them to the user. The product value also stands in the fact that it can provide a dynamic content to the consumers, information tailored to a certain user (e.g., a set of diseases that the consumer is interested in).

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 [25], that we also use in our research.

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 to test its performance. The objectives are to build models that can later be deployed on other test sets with high performance.

A. Machine Learning Overview

To solve a problem on a computer, we need an algorithm. An algorithm is a sequence of instructions that should be carried out to transform the input to output. For example, one can devise an algorithm for sorting. The input is a set of numbers and the output is their ordered list. For the same task, there may be various algorithms and we may be interested in finding the most efficient one, requiring the least number of instructions or memory or both. For some tasks, however, we do not have an algorithm—for example, to tell spam emails from legitimate emails. We know what the input is: an email document that in the

simplest case is a file of characters. We know what the output should be: a yes/no output indicating whether the message is spam or not. We do not know how to transform the input to the output. What can be considered spam changes in time and from individual to individual.

With advances in computer technology, we currently have the ability to store and process large amounts of data, as well as to access it from physically distant locations over a computer network. Most data acquisition devices are digital now and record reliable data. We may not be able to identify the process completely, but we believe we can construct *a good and useful approximation*. That approximation may not explain everything, but may still be able to account for some part of the data. We believe that though identifying the complete process may not be possible, we can still detect certain patterns or regularities. This is the niche of machine learning. Application of machine learning methods to large databases is called *data mining*. But machine learning is not just a database problem; it is also a part of artificial intelligence. To be intelligent, a system that is in a changing environment should have the ability to learn. If the system can learn and adapt to such changes, the system designer need not foresee and provide solutions for all possible situations.

Machine learning is programming computers to optimize a performance criterion using example data or past experience. We have a model defined up to some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or past experience. The model may be *predictive* to make predictions in the future, or *descriptive* to gain knowledge from data, or both.

In *medical diagnosis*, the inputs are the relevant information we have about the patient and the classes are the illnesses. The inputs contain the patient's age, gender, past medical history, and current symptoms. Some tests may not have been applied to the patient, and thus these inputs would be missing. Tests take time, may be costly, and may inconvience the patient so we do not want to apply them unless we believe that they will give us valuable information. In the case of a medical diagnosis, a wrong decision may lead to a wrong or no treatment, and in cases of doubt it is preferable that the classifier reject and defer decision to a human expert

Kernel machines are maximum margin methods that allow the model to be written as a sum of the influences of a subset of the training instances. These influences are given by application-specific similarity kernels, and we discuss "kernelized" classification, regression, outlier detection, and dimensionality reduction, as well as how to choose and use kernels

For our experiments, the text was obtained from MEDLINE 20012. An annotator with biology expertise considered the titles and abstracts separately and labeled the sentences (both roles and relations) based solely on the content of the individual sentences. Seven possible types of relationships between TREATMENT and DISEASE were identified. It shows, for each relation, its definition, one example sentence and the number of sentences found containing it. The results reported in this paper were obtained with the following features: the word itself, its part, of speech from the Brill tagger (Brill, 1995), the phrase constituent the word belongs to, obtained by flattening the output of a parser (Collins, 1996), and the word's MeSH ID (if available). In addition, we identified the sub-hierarchies of MeSH that tend to correspond to treatments and diseases, and convert these into a tri-valued attribute indicating one of: disease, treatment or neither. Finally, we included orthographic features such as 'is the word a number', 'only part of the word is a number', 'first letter is capitalized', and 'all letters are capitalized'. In we analyze the impact of these features.

B. Natural Language Processing Overview

Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.

The goal of NLP as stated above is "to accomplish human-like language processing". The choice of the word 'processing' is very deliberate, and should not be replaced with 'understanding'. For although the field of NLP was originally referred to as Natural Language Understanding (NLU) in the early days of AI, it is well agreed today that while the goal of NLP is true NLU, that goal has not yet been accomplished. A full NLU System would be able to:

1. Paraphrase an input text
2. Translate the text into another language
3. Answer questions about the contents of the text
4. Draw inferences from the text

Naturally occurring texts' can be of any language, mode, genre, etc. The texts can be oral or written. The only requirement is that they be in a language used by humans to communicate to one another. Also, the text being analyzed should not be specifically constructed for the purpose of the analysis, but rather that the text be gathered from actual usage. It runs into many stages, namely tokenization, lexical analysis, syntactic analysis, semantic analysis, and pragmatic analysis. Syntactic analysis provides an order and structure of each sentence in the text. Semantic analysis is to find the literal meaning, and pragmatic analysis is to determine the meaning of the text in context. These major tasks are further broken down into, parsing and so on.

For the work presented in this paper, the data sets contain sentences that are annotated with the appropriate information. Unlike in the work of Rosario and Hearst [25], in our research, the annotations of the data set are used to create a different task (task 1). It identifies informative sentences that contain information about diseases and treatments and semantic relations between them, versus noninformative sentences. This allows us to see how well ML techniques can cope with the task of identifying informative sentences, or in other words, how well they can weed out sentences that are not relevant to medical diseases and treatments. In the first task, the data sets are annotated with the following information: a label indicating that the sentence is informative, i.e., containing disease-treatment information, or a label indicating that the sentence is not informative. Table 2 gives an example of labeled sentences. In the second task, the sentences have annotation information that states if the relation that exists in a sentence between the disease and treatment is Cure, Prevent, or Side Effect. These are the relations that are more represented in the original data set and also needed for our future research. We would like to focus on a few relations of interest and try to identify what predictive model and representation technique bring the best results.

C. 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.

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.

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. 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

V. DISCUSSION

This section discusses the results we obtained for the two tasks in this study. For the first task, the one of identifying informative sentences, the results show that probabilistic models based on Naı¨ve Bayes formula, obtain good results. The fact that the SVM classifier performs well shows that the current discoveries are in line with the literature. These two classifiers have always been shown to perform well on text classification tasks. Even though the independence of features is violated when using Naı¨ve Bayes classifiers, they still perform very well.

In NLP and ML community, BOW is a representation technique that even though it is simplistic, most of the times it is really hard to outperform. As shown in Fig. 9, the results obtained with this representation are among the best one, but for both tasks, we outperform it when we combine it with more structured information such as medical and biomedical concepts.
The first bars of results are obtained with the best model for each of the eight relations (e.g., for Cure, the representation that obtains the best results is reported, a representation that can be different from the one for another relation; the label of each set of bars describes the representation); the second bars of results report the model that obtains the best accuracy over all relations (one representation and one classification algorithm are reported for all relations—CNB), and the third bars of results represent the previous results obtained by Rosario and Hearst [25].

VI. CONCLUSION

The conclusions of our study suggest that domain-specific knowledge improves the results. Probabilistic models are stable and reliable for tasks performed on short texts in the medical domain. The representation techniques influence the results of the ML algorithms, but more informative representations are the ones that consistently obtain the best results.

The first task that we tackle in this paper is a task that has applications in information retrieval, information extraction, and text summarization. We identify potential improvements in results when more information is brought in the representation technique for the task of classifying short medical texts. We show that the simple BOW approach, well known to give reliable results on text classification tasks, can be significantly outperformed when adding more complex and structured information from various ontologies.

The second task that we address can be viewed as a task that could benefit from solving the first task first. In this study, we have focused on three semantic relations between diseases and treatments. Our work shows that the best results are obtained when the classifier is not overwhelmed by sentences that are not related to the task. Also, to perform a triage of the sentences (task 1) fora relation classification task is an important step. As future work, we would like to extend the experimental methodology when the first setting is applied for the second task, to use additional sources of information as representation techniques, and to focus more on ways to integrate the research discoveries in a framework to be deployed to consumers.

REFERENCES

[1] R. Bunescu and R. Mooney, “A Shortest Path Dependency Kernel for Relation Extraction,” Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing (HLT/ EMNLP), pp. 724-731, 2005.

[2] R. Bunescu, R. Mooney, Y. Weiss, B. Schoolk¨ lkopf, and J. Platt, “Subsequence Kernels for Relation Extraction,” Advances in Neural Information Processing Systems, vol. 18, pp. 171-178, 2006.

[3] A.M. Cohen and W.R. Hersh, and R.T. Bhupatiraju, “Feature Generation, Feature Selection, Classifiers, and Conceptual Drift for Biomedical Document Triage,” Proc. 13th Text Retrieval Conf. (TREC), 2004.

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

[5] I. Donaldson et al., “PreBIND and Textomy: Mining the Biomedical Literature for Protein-Protein Interactions Using a Support Vector Machine,” BMC Bioinformatics, vol. 4, 2003.

[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] O. Frunza and D. Inkpen, “Textual Information in Predicting Functional Properties of the Genes,” Proc. Workshop Current Trends in Biomedical Natural Language Processing (BioNLP) in conjunction with Assoc. for Computational Linguistics (ACL ’08), 2008.

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

[9] C. Giuliano, L. Alberto, and R. Lorenza, “Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature,” Proc. 11th Conf. European Chapter of the Assoc. for Computational Linguistics, 2006.

[10] J. Ginsberg, H. Mohebbi Matthew, S.P. Rajan, B. Lynnette, S.S. Mark, and L. Brilliant, "Detecting Influenza Epidemics Using Search Engine Query Data," Nature, vol. 457, pp. 1012-1014, Feb. 2009.

[11] M. Goadrich, L. Oliphant, and J. Shavlik, "Learning Ensembles of First-Order Clauses for Recall-Precision Curves: A Case Study in Biomedical Information Extraction," Proc. 14th Int'l Conf. Inductive Logic Programming, 2004.

[12] L. Hunter and K.B. Cohen, "Biomedical Language Processing: What's beyond PubMed?" Molecular Cell, vol. 21-5, pp. 589-594,

[13] L. Hunter, Z. Lu, J. Firby, W.A. Baumgartner Jr., H.L. Johnson, P.V. Ogren, and K.B. Cohen, "OpenDMAP: An Open Source, Ontology-Driven Concept Analysis Engine, with Applications to Capturing Knowledge Regarding Protein Transport, Protein Interactions and Cell-Type-Specific Gene Expression," BMC Bioinformatics, vol. 9, article no. 78, Jan. 2008.

[14] T.K. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig, "A Literature Network of Human Genes for High-Throughput Analysis of Gene Expression," Nature Genetics, vol. 28, no. 1, pp. 21-28, 2001.

[15] R. Kohavi and F. Provost, "Glossary of Terms," Machine Learning, Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process, vol. 30, pp. 271-274, 1998.

[16] G. Leroy, H.C. Chen, and J.D. Martinez, "A Shallow Parser Based on Closed-Class Words to Capture Relations in Biomedical Text," J. Biomedical Informatics, vol. 36, no. 3, pp. 145-158, 2003.

[17] J. Li, Z. Zhang, X. Li, and H. Chen, "Kernel-Based Learning for Biomedical Relation Extraction," J. Am. Soc. Information Science and Technology, vol. 59, no. 5, pp. 756-769, 2008.

[18] T. Mitsumori, M. Murata, Y. Fukuda, K. Doi, and H. Doi, "Extracting Protein-Protein Interaction Information from Biomedical Text with SVM," IEICE Trans. Information and Systems, vol. E89D, no. 8, pp. 2464-2466, 2006.

[19] M. Yusuke, S. Kenji, S. Rune, M. Takuya, and T. Jun'ichi, "Evaluating Contributions of Natural Language Parsers to Protein-Protein Interaction Extraction," Bioinformatics, vol. 25, pp. 394-400, 2009.

[20] S. Novichkova, S. Egorov, and N. Daraselia, "MedScan, A Natural Language Processing Engine for MEDLINE Abstracts," Bioinformatics, vol. 19, no. 13, pp. 1699-1706, 2003.

[21] M. Ould Abdel Vetah, C. Ne´dellec, P. Bessie`res, F. Caropreso, A.-P. Manine, and S. Matwin, "Sentence Categorization in Genomics Bibliography: A Naive Bayes Approach," Actes de la Journe´e Informatique et Transcriptome, J.-F. Boulicaut and M. Gandrillon, eds., Mai 2003.

[22] J. Pustejovsky, J. Capstan 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.

AUTHORS

First Author – R.Buvaneshwari**,** PG Student, Department of Information Technology, Periyar maniammai University, Vallam, Thanjavur, India, Email id - buvi.sumathi@gmail.com

Second Author – K.Lavanya, PG Student, Department of Information Technology, Periyar maniammai University, Vallam, Thanjavur, India, Email id – lavanya62.cse@gmail.com

Third Author – R.Vanitha, Asst.Professor, Department of Information Technology, Periyar maniammai University, Vallam, Thanjavur, India , Email id - *[3]Vani_34@rediffmail.com*