Diagnosticating and Propagating Health Maintenance Information Using Machine Learning Based Methodology

P.Deepa, M.Baskar

Abstract:-The Machine Learning field has gained its thrust in almost any domain of research and just recently has become a reliable tool in the medical system. The experiential 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. Machine Learning is a tool by which computer-based systems can be integrated in the healthcare field in order to get a better, well-organized medical care. It describes a ML based methodology for building an application that is capable of identifying and disseminating healthcare information. It extracts sentences from published medical papers that mention diseases and treatments, and identifies semantic relations that exist between diseases and treatments. The traditional healthcare system is also becoming one that embraces the Internet and the electronic world. Electronic Health Records (hereafter, EHR) are becoming the standard in the healthcare domain. 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 health care. In order to embrace the views that the EHR system has need better, faster, and more reliable access to information. The proposed system is to show Natural Language Processing (NLP) and Machine Learning (ML) techniques and what classification algorithms are suitable to use for identifying and classifying relevant medical information in short texts. It recognize the fact that able to identifying reliable information in the medical system stand as construction blocks for a healthcare system that is up-to-date with the latest discoveries. In this examines, mainly focus on diseases and treatment information, and the relation that exists between these two entities. The main work is to automatically identifying sentences published in medical papers as containing or not information about diseases and treatments, and also identifying semantic relations that exist between diseases and treatments.

Index Terms—Healthcare, machine learning, natural language processing

1. INTRODUCTION

1.1 Data Mining-Overview

Data mining refers to extracting knowledge or mining interesting knowledge from large amount of data stored either in databases or other information repositories, a standard definition for data mining is the non-trivial extraction of implicit, previously unknown and potentially useful knowledge from data. The alternative name of the term Data Mining is knowledge extraction, business intelligence, data archaeology, data/pattern analysis, information harvesting, software and even data dredging, and describes the concept of knowledge discovery in databases.

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. Data mining software is one of the numbers of analytical tools for analyzing data. It allows users to analyze data from different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Knowledge Discovery in Databases (KDD) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in large data collections. The most important step within the process of KDD is data mining which is concerned with the extraction of the valid patterns. KDD is necessary to analyze the steady growing amount of data caused by the enhanced performance of modern computer systems. However, with the growing amount of data the complexity of data objects increases as well. Modern methods of KDD should therefore examine more complex objects than

ISSN: 2278 – 7798

simple feature vectors to solve real-world KDD applications adequately.

Data mining software analyzes relationships and patterns in stored transaction data based on open-end user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

Classes: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.

Clusters: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.

Associations: Association mining aims to extract interesting correlation, frequent patterns, associations or casual structures among set of items or objects in transaction databases. Data can be mined to identify associations. The beer-diaper example is an example of associative mining. Association rules are widely used in various areas such as telecommunication networks, market and risk management, clustering, classification, etc.

Sequential patterns: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

Classification: Classification is to learn the function that maps a data item into one of several predefined classes.

Data mining consists of five major elements:

- Extract, transform, and load transaction data into the data warehouse system
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

Data mining derives its name from the similarities between searching for valuable information in a large database and mining a mountain for a vein of valuable one. Both processes require either sifting through an immense amount of material, or intelligently probing it to find where the value resides.

1.1.1 Data Preprocessing

The importance of data preprocessing concept takes place if the quality of data is not good, mining poor results

Data warehouse needs consistent integration of quality data. Quality decisions must be based on quality data. Data extraction, cleaning, and transformation comprise the majority of the work of building a data warehouse.
Major task in data mining as follows:

Data cleaning: Fill the missing values, smooth noisy data, remove outliers.

Data Integration: Integration of multiple databases, data cubes, or flat files.

Data Transformation: Normalization and aggregation

Data Reduction: Obtains reduce data size and produce same analytical results.

Data Discretization: Part of data reduction, especially for numerical data.

Incorrect attribute values may due to

- Faulty data collection instruments
- Data entry problems
- Data transmission problems
- Technology limitations
- Inconsistency in naming conventions

1.1.2 Cluster Analysis

Cluster is a collection of data objects, it groups the similar data objects within the same cluster and dissimilar objects in other clusters. Cluster analysis is finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters.

A good clustering method will produce high quality clusters with high intra-class similarity which is similar to one another within the same cluster, low inter-class similarity which is dissimilar to the objects in other clusters are the quality of a clustering method is also measured by its ability to discover some of the hidden patterns, distances are normally used to measure the

ISSN: 2278 – 7798

similarity or dissimilarity between two data objects.

In general, the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the interclass similarity and minimizing the interclass similarity.

2. RELATED WORK

2.1 Analysis of Related Works

To better understand of improving reliability, better and feasibility, it is useful to review and examine the existing research works in literature. Therefore, recent approaches and methodologies used for improving reliability and feasibility have been discussed.

Shawe-Taylor and Cristianini (2005) discusses about the relation extraction, shallow linguistic processing did not show their potential when an explicit computation of the feature map becomes computationally infeasible, due to the high or even infinite dimension of the feature space. The main reason concerns the fact that syntactic parsing is not always robust enough to deal with real-world sentences. This may prevent approaches based on syntactic features from producing any result.

Another related issue concerns the fact that parsers are available only for few languages and may not produce reliable results when used on domain specific texts (as is the case of the biomedical literature). For this reason, kernels have been recently used to develop innovative approaches to relation extraction based on syntactic information. The convolution kernel technique identifies sentences describing interactions with a precision of more yielding significant improvements over machine learning techniques.

Bunescu and Mooney (2006) discuss about the sentence identification, simple linguistic features provides two methods to extract the sentence: the global context where entities appear and their local context. The whole sentence where the entities appear (global context) is used to discover the presence of a relation between two entities around the entities (local context) provide useful clues to identify the roles of the entities within a relation. The proposed approach, perform the extraction task in a single step via a combined kernel while in the previous steps used two separate classifiers to identify entities and relations and their output is later combined with a probabilistic global inference.

Ray and Craven (2007) discuss about the information extraction, accuracy level of finding relation between the sentence is not efficient it examines a key of natural language processing which is syntactic and semantic parsers on natural language text from different domains limit the extent to which syntactic and semantic information can be used in real systems.

Richards and Mooney (2008) discuss about the protein to protein interaction, over the past few years, a multitude of high throughput methods to detect protein interactions have been developed. Consolidating the known list of protein-protein interactions will provide researchers with a powerful tool that will greatly enhance their understanding of these relationships on a genomic scale. This is to efficiently mine protein interactions with high precision.

There have been several attempts to develop databases of interacting proteins, and the supporting metadata that describes an interaction. Currently, there are several manually created databases like DIP, BIND and MINT. While these databases promise a high degree of accuracy and fast.

Culotta and Sorensen (2009) discuss about the text classification and regression, SVMs are able to solve a multitude of classification problems by using domain-specific and cost-sensitive kernel functions. The proposed approach is to use a tree kernel to efficiently determine the similarity between the syntactic parse trees of sentences. As a result, able to exploit the syntactic structure of natural language text.

Zelenko et al (2010) discuss about the search space problem, it require a method to prune the search space, so that can apply more accurate and complex machine learning techniques on a reduced candidate set. By using SVMs with a Bag-of-Words approach to build a classifier model which successfully distinguishes those abstracts which may contain a protein interaction. The Bag-of-words technique is a simple approach that relies on word-frequency information to classify text documents.

ISSN: 2278 – 7798

Esposito and Malerba (2011) discuss about the entity recognition, maximum entropy models is not relatively perform entity recognition and relation discrimination. The proposed work introduced here is Hidden Markov Models to perform these techniques are based on words in context, part of speech information, phrases, and a medical lexical ontology. Generates improved results with less annotated data.

3. METHODOLOGY

3.1 Machine Learning

The propose system approach, this work is to show what Natural Language Processing (NLP) and Machine Learning (ML) techniques what demonstration of information and what classification algorithms are suitable to use for identifying and classifying relevant medical information in short texts. It recognize the fact that tools able of identifying reliable information in the medical domain stand as construction blocks for a healthcare system that is up-to-date with the latest discoveries. In this examine, mainly focus on diseases and treatment information, and the relation that exists between these two entities. This 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 ML-based methodology for building an application that is capable of identifying and disseminating healthcare information. It extracts sentences from published medical papers that mention diseases and treatments, and identifies semantic relations that exist between diseases and treatments. In addition to more methodological settings in which try to find the potential value of other types of representations, and it would like to focus on source data that comes from the web. Identifying and classifying medical-related information on the web is a challenge that can bring valuable information to the research community and also to the end user.

The identification of required data and domain knowledge requires the collaboration with a domain expert and is an important step of the process of applying ML to real-world problems. Only recently, the related issues of feature selection and, more generally, data preprocessing have been more systematically investigated in ML. Data preprocessing is still considered a step of the knowledge discovery process and confined to data cleaning, simple data transformations (e.g., summarization) and validation. On the contrary, many studies in computer vision and pattern recognition focused on the problems of feature extraction and selection. Their properties have been well investigated and available tools make their use simple and efficient.

Machine Learning Algorithm

Input: {request}, {advertisement}
Output: {connection}
Processing:
Initialize {connection} to Empty
For each reqi ∈ {request}
For each advj ∈ {advertisement}
If advj..availability is true **AND**
reqi .CPUSpeed is less than or equal to advj .CPUSpeed **AND**
reqi .Memory is less than or equal to advj .Memory **AND**
reqi .OS is equal to advj .OS **AND**
reqi .Disk capacity is less than or equal to advj .Disk capacity **AND**
reqi .timeslot ∩ advj .timeslot is not empty **AND**
reqi .max_acceptable_price is less than advj .min_acceptable_price
Then
Add reqi and advj into {connection}
End-if
End-for
End-for

3.2 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 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, or laypeople who want to be in charge of their health by reading the latest life science published articles related to their interests. The

ISSN: 2278 – 7798

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 can be developed and sold by companies that do research in Healthcare Informatics, Natural Language Processing, and Machine Learning, and companies that develop tools like Microsoft Health Vault. The value of the product from an e-commerce point of view stands in the fact that it can be used in marketing strategies to show that the information that is presented is trustful (Medline articles) and that the results are the latest discoveries.

For any type of business, the trust and interest of customers are the key success factors. Consumers are looking to buy or use products that satisfy their needs and gain their trust and confidence. Healthcare products are probably the most sensitive to the trust and confidence of consumers. 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 diseases-treatment relations in the sentences already selected as being informative (e.g., task 1 is applied first). Here mainly focus on three relations: Cure, Prevent, and Side Effect, a subset of the eight relations that the corpus is annotated with.

Normally 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. The task of identifying the three semantic relations is addressed in two ways: 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 labelled either with the relation in question or with non relevant information. 2. One model is built, to distinguish the three relations in a three-class classification task where each sentence is labelled with one of the semantic relations

4. EXPERIMENTS AND RESULTS

EXPERIMENTS AND RESULTS

This section discusses the evaluation measures and presents the results of the two tasks using the methodology described above. Perform Sentence Identifying as task 1 and Relation Identification as task 2.

4.1 Evaluation Measures

The most common used evaluation measures in the ML settings are: accuracy, precision, recall, and F-measure. All these measures are computed form a confusion matrix that contains information about the actual classes, the true classes and the classes predicted by the classifier. The test set on which the models are evaluated contain the true classes and the evaluation tries to identify how many of the true classes were predicted by the model classifier. In the ML settings, special attention needs to be directed to the evaluation measures that are used. For data sets that are highly imbalanced standard evaluation measures like accuracy are not suitable. Because the data sets are imbalanced, user chose to report in addition to accuracy, the macro averaged F-measure. Here decided to report macro and not micro averaged F-measure because the macro measure is not influenced by the majority class, as the micro measure.

The macro measure better focuses on the performance the classifier has on the minority classes. The formulas for the evaluation measures are: Accuracy ¼ the total number of correctly classified instances, Recall ¼ the ratio of correctly classified positive instances to the total number of positives. This evaluation measure is known to the medical research community as sensitivity. Precision ¼ the ratio of correctly classified positive instances to the total number of classified as positive. F-measure ¼ the harmonic mean between precision and recall.

4.2 Results for Task of Identifying Informative Sentences

This section presents the results for the first task, the one of identifying whether sentences are informative, i.e., containing information about diseases and treatments, or not. The ML settings

ISSN: 2278 – 7798

are created for a two-class classification task and the representations, while the baseline on which need to improve is given by the results of a classifier that always predicts the majority class.

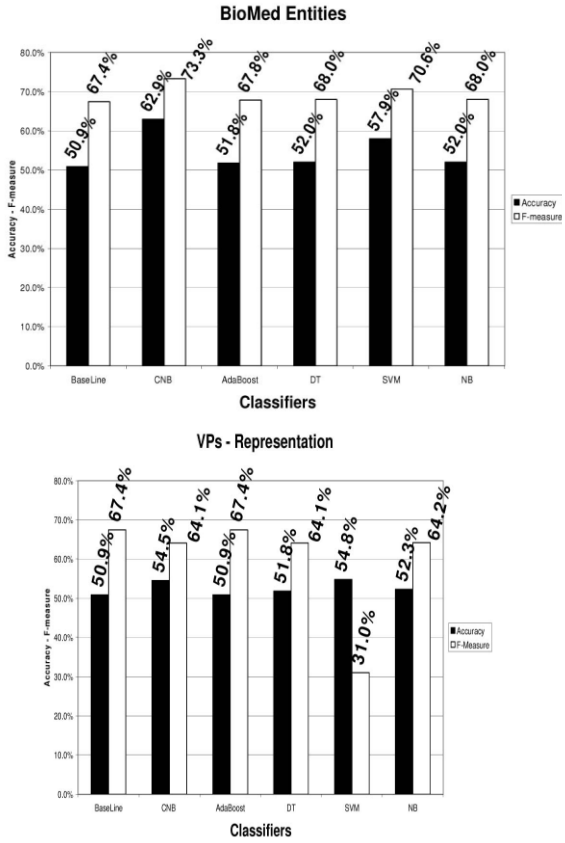

Fig 4.2.1 Accuracy and F-measure results using VP

It presents the results obtained when using as representation features verb-phrases identified by the Genia tagger. When using this representation, the results are close to baseline. The reason why this happens for all algorithms that we use is the fact that the texts are short and the selected features are not well represented in an instance. We have a data sparseness problem: it is the case when a lot of features have value 0 for a particular instance.

Fig 4.2.2 Accuracy and F-measure results using Biomedical

It presents the results obtained using as representation features noun-phrases selected by the Genia tagger. Compared to previous results, we can observe a slight improvement in both accuracy and F-measure. The best results are obtained by the CNB classifier. We believe that the slight improvement is due to a reduction of the sparseness problem: noun-phrases are more frequently present in short texts than verb-phrases.

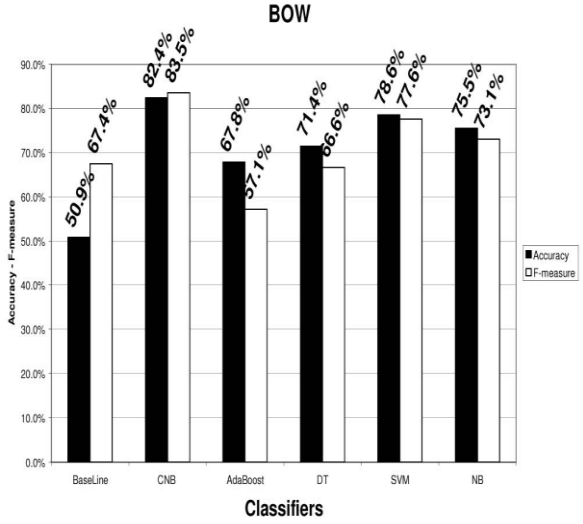

Fig 4.2.3 Accuracy and F-measure results using BOW

The bag-of-words representation technique is known in the literature to be one that is hard to beat. Even though is not a very sophisticated method—it contains only the words in context; it is one that often obtains good results. In this experiment, the BOW representation (Fig. 4.2.3) obtains the best results between all the representation techniques.

4.3 Results for Task of Identifying Semantic Relations

The focus for the second task is to automatically identify which sentences contain information for the three semantic relations: Cure, Prevent, and Side Effect. The reported results are based on similar settings to the ones used for the previous task. Since imbalanced data sets are used for this task, the evaluation measures that are going to report is the F-measure. Due to space issues, we are going to present the best results obtained for all settings. The best results are chosen from all the representation techniques and all classification algorithms that we also used for the first task. The labels on the x-axis stand for the name of the semantic relation, the representation technique, and the classification algorithm used.

ISSN: 2278 – 7798

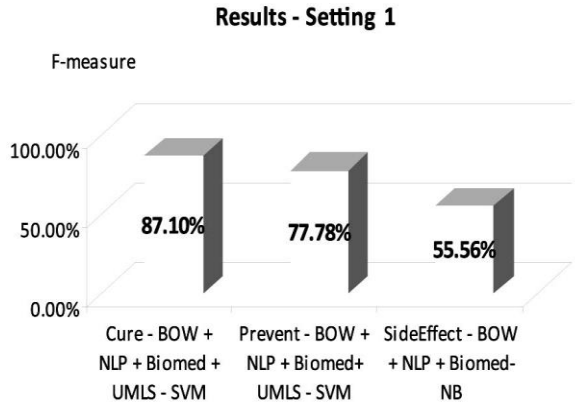

Fig 4.3.1 Results for Setting 1.

On the x-axis, present for each relation the best F-measure result, the representation technique, and the classifier that obtained the result. For example, for the Cure relation, the combination of BOW features, noun-phrases and verb phrases, biomedical and UMLS concepts, with SVM as a classifier, obtained the 87.10 percent result for F-measure. SVM and NB with rich feature representations are the setups that obtained the best results.

For example, for the Cure relation, the combination of BOW features, noun-phrases and verb phrases, biomedical and UMLS concepts, with SVM as a classifier, obtained the 87.10 percent result for F-measure. SVM and NB with rich feature representations are the

5. CONCLUSION

The conclusions 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 is to tackle in this paper is a task that has applications in information retrieval, information extraction, and text summarization.

Identify potential improvements in results when more information is brought in the representation technique for the task of classifying short medical texts. Here 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 ontology's. The second task that to address can be viewed as a task that could benefit from solving the first task first.

In this study, it mainly focused on three semantic relations between diseases and treatments. The 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) for a relation classification task is an important step. In Setting 1, it included the sentences that did not contain any of the three relations in question and the results were lower than the one when we used models trained only on sentences containing the three relations of interest. These discoveries validate the fact that it is crucial to have the first step to weed out uninformative sentences, before looking deeper into classifying them. Similar findings and conclusions can be made for the representation and classification techniques for task 2.

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

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

[4] 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.

[5] 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.

[6] 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.

[7] J. Ginsberg, H. Mohebbi Matthew, S.P. Rajan, B. Lynnette, S.S. Mark, and L. Brilliant, "Detecting

ISSN: 2278 – 7798

Influenza Epidemics Using Search Engine Query Data," Nature, vol. 457, pp. 1012-1014, Feb. 2009.
[8] 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.
[9] 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
[10] 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..
[11] 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.
[12] 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.

P.Deepa, currently pursuing M.Phil Computer Science under one of the college affiliated to Periyar University. I also received my B.Sc and M.Sc degrees from the affiliated Colleges under Anna University.

I have done internship during my Post Graduate.

Mr M.Baskar, currently acting as Assistant Professor, Department of Computer Science in Vivekanandha College for Women. He doing research in Data Mining.

ISSN: 2278 – 7798