Applications of Text Classification using Text Mining

Mrs. Manisha Pravin Mali[1], Dr. Mohammad Atique[2]
[1]Dept. of Computer Engg., Vishwakarma Institute of Information Technology, Pune, India
[2]Dept. of Computer Science, Sant Gadge Baba Amravati University, Amravati, India

***Abstract*— Text mining is a technology to discover patterns, trends and knowledge which is previously unknown, semi-automatically from huge collections of unstructured text. Text classification in text mining is a supervised learning process, which aims to assign a document to one or more predefined categories based on its content. Text classification using text mining helps us to analyse large amount of digital data. In this paper, some of the applications are discussed which are belonging to areas like business, medicine, law and society.**

Keywords— **Text classification, Text mining, Business, Medicine, Law, Society**

I. INTRODUCTION

In today's world, huge amount of information is available in digital form. Most of these information is in unstructured form. So we need a proper tool which can help us to convert this information into knowledge. Text mining is the tool to extract useful information from unstructured textual data through the identification and exploration of interesting and useful patterns.There are some text mining functions which give solutions for common information management problems. Text mining function involves searching, classification, clustering, summarization, information monitor and information extraction. Searching provide a search interface to an information collection and has become common and almost essential for usability. Information Extraction can converts unstructured text into a form that can be loaded into a tabular database. An information monitor periodically collects, organizes, and analyzes articles from multiple sources. When the description of categories is unknown, clustering is used. The purpose of summarization is to accurately summarize i.e. convey the essence of a document with as little text as possible. There are many applications which needs to classify the textual data. Text classification is the task of automatically classifying a set of documents into a predefined classes[3]. Text classification process consists of steps like pre-processing, indexing, dimensionality reduction & feature selection, and classification. Some of the well-known techniques for text classifications are Decision tree, Rule based classifier, Nearest-Neighbour classifier, Bayesian classifier, Support Vector Machine and Artificial Neural Network[5]. Also research is going on various techniques belonging to soft computing to improve accuracy and efficiency of classification[4].

This paper provides an architecture of text mining applications in sections II. Various Text classification applications have been discussed in section III.

II. AN ARCHITECTURE FOR TEXT MINING APPLICATIONS

In today's world, huge amount of information is available in digital format. Text mining tools can examine through these vast repositories of online or digital information to find relevant data. This information can be retrieve from more than one source like email, the Web, newsgroups, intranets and so on. Each source may have a separate practice and interface. The end user has to process information from each source separately and manually synthesizes collected information. The types of documents of collected information include emails, Web pages, audio transcripts, articles, formatted files, and postings on a newsgroup. Relationship may exist within this document collection, also documents may scattered over multiple directories. Initially, raw documents/data gets collected from various sources for different application domains/areas. Then application building gets started. The text within documents is often formatted using a tagging language or a proprietary format. Plain text converter i.e. pre-processing step, converts unstructured data into structured data. Each converter is unique and converts a particular data format to structured data or plain text. In this step, token formation is done with tokenization process. A token is a word, number, punctuation mark, or any other sequence of characters that should be treated as a single unit. After tokenization, stop word removing process followed by stemming process, is done. After this, we get the set of unique words occurring in a document or the collection weighted by their importance in the document which is a reasonable representation. This process is also known as vector generation. The term *vector is used* for such a representation since each token has a dimension (direction) and has some weight (magnitude). After vector generation, document indexing is done. Indexed documents then stored in a text accumulator. The *indexing* is a compressed representation of a collection of documents used to retrieve documents with a set of words. Then various text mining techniques like categorization, clustering, summarization, and so on, can be apply for specific knowledge extraction. (Fig. 1)

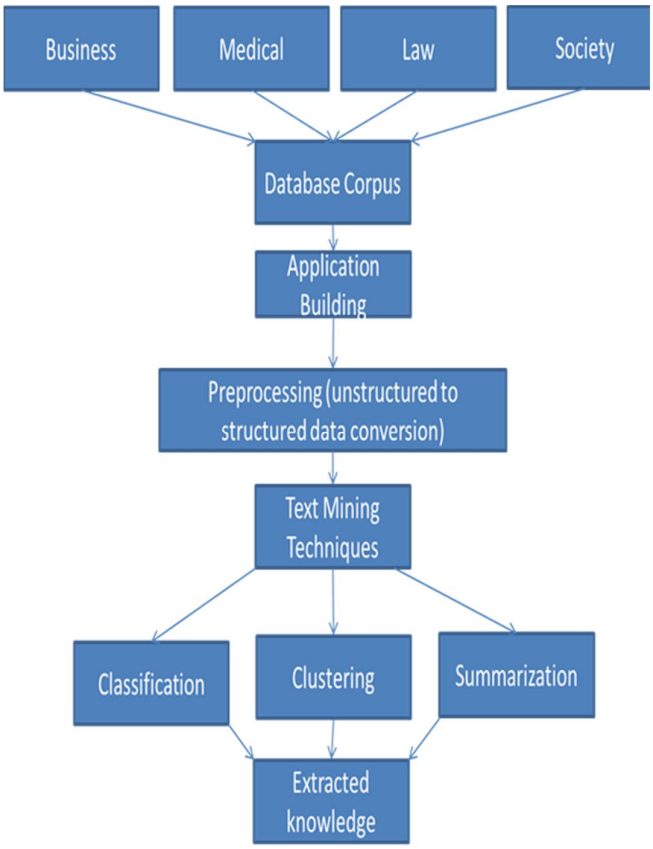

FIGURE 1 Architecture of text mining application building

III. Applications

We get benefits of text mining in our daily life often unknowingly. For example, the emails sent to us may be filtered through a text mining tool before being delivered to us. Like this there are many applications of text mining in various areas. In this section we will go through some applications of text mining. Broadly these applications are categories in four main areas: business, medicine, law, and society, but not limited to these areas.

.

A. Business

Corporations are some of the biggest consumers and producers of information. Huge amount of stored information lie unused. Below applications show how some of this information is used.

Most companies are constantly evolving to compete better. Before making changes, it would be useful to judge potential employees or customer reactions (opinion). This is done by different surveys like quantitative or descriptive surveys. In case of descriptive survey, questions accept natural language answers, which are very difficult to process. For this "Summarization" can help. In this sentiment analysis is used to categorize documents based on the nature of the text. Sentiment analysis identifies positive and negative opinions and emotions based on the occurrence of positive/negative words defined in a vocabulary.

To monitor competitor company's products, announcements, and developments is routine practice in industry. Doing this manually is very slow process, as sites for large companies may have several hundred Web pages. An automated approach to periodically download and monitor a competitor's Web pages is more efficient. For this task, we can build a smart crawler to search the Web. Some Web sites publish information frequently and the contents may change frequently. Announcements are made, products are released, or reviews are generated. Visiting such sites daily to check what's new is uninteresting and tedious. An information monitor can scan such sites, identify common articles, and summarize the information from multiple news sites in a single Web page. News articles are typically well written and somewhat easier to extract information from than a general Web page. Information about companies making investments, mergers, and related financial information can be captured in a template. A monitor can also extract the names of people or organizations mentioned in context with the help of "Information Extraction".

Companies safeguard their intellectual property carefully. It can be the source of a company's profits, providing a good reason to keep it secure. One way of protecting Intellectual Property is by filing patents. Big companies may file hundreds or even thousands of patents yearly. The main purpose of a patent is to prevent a competitor from getting an idea that a company may have thoroughly developed. With thousands of patents on file, it is difficult to check the innumerable products launched yearly for infringement. Manually matching thousands of products against thousands of patents is difficult and may cause some error. Text descriptions of products and patents can be automatically compared. Computations where the similarity is above a threshold can be considered for verification. For this document Clustering is useful.

To judge product's success, customer satisfaction is one of the important criteria. For these customers feedback is taken in various forms like hard coded or through emails. These collected feedback needs to be classify to understand customers satisfaction level. Also companies want good relationship with customer for further business. For this companies have their own Customer relationship management (CRM) cell[2]. CRM covers a set of processes to allowing systems supporting a business strategy to build long term and profitable relationships with explicit task of handling complaints arriving in a number of languages customers. CRM consists of various dimensions like Customer identification, attraction, preservation and development. Email communication from customer plays an important role to work in various dimensions of CRM. So companies need a system to classification accurate and automatic. Accurately categorizing email can be achieved with "Text Categorization/Classification". Multinational companies

market their products in many countries. Customer feedback may be received in multiple languages. Automatic language detection and machine translation can simplify the. In this cross language information retrieval is also useful.

Online recruitment is a huge industry. They collect resumes of candidate through emails or through a company's web site. Some companies use a special web site that specializes in matching candidates and requirements. A company may get hundreds or thousands of resumes, depending on the market and the demand. Handling these numbers of resumes is laborious, and, in some cases, the law does not allow employers to delete email responses and each applicant's resume must be filed. Preliminary information like educational qualifications, work experience, rate of job changes, job designations, job awards and other personal information can be used in filtering resumes in first step. For this automatic information extractor is useful. In second step weight or rank may be assigned to resumes depending on different attribute values. An ideal resume of a candidate for a required vacancy with the flawless educational qualifications and work experience can be built. Attributes of submitted resumes and ideal resume can be compared to decide whether to call a candidate for further steps or not. This can reduce a lots of human efforts required against resume scrutiny to manage e-recruitment process.

In a knowledge-rich society, there is too much information available for one person to absorb. When we want information on specialized topic, we need to consult experts for specialized expertise. Experts in an organization are usually known by the information published within an organization, recognition from external sources, and referral by examines. The flow conformation within an organization can be monitored, collected, and analysed to locate experts. This operation can occur transparently and does not require additional input from an employee. Information in the form of emails, presentations, Web pages, and formatted documents on an intranet provides a basis for identifying experts.

B. Medicine

Practitioners of medicine generate and consume large volumes of information. In medicine, studies, research reports, clinical trials, hospital records, and doctors' notes constitute sources of information. Most of this information is in the form of text. Also there is a lot of interest in the research of genes and proteins like other medical domains. To read and analyse all the information manually is difficult. So tools to extract information from such a large databases are required. These tools are built exclusively to mine medical or scientific literature or information. Some tools capture the interaction between cells, molecules, and proteins, and others extract biological facts from articles. Thousands of these facts can be automatically analysed for similarities or relationships [1].

C. Law

Huge volumes of legal text information are generated on a daily basis. Legal information includes court transcripts, statements and affidavits taken, written motions, verdicts, and lawsuit declarations. Organizing this information for easy access has been a problem for decades. Buried in legal information are interpretations of the law in a case, relationships between police reports and written statements, and trends or patterns in society. Using a search engine alone to extract this information is not the preferred method. Sometimes the legal researcher may not know the keywords or specific terminology to precisely describe the information needed. So, automated tools that can extract and summarize the information have many benefits over a search engine.

D. Society

There are many text mining applications which are specifically built for society. Main purpose behind this is to collect and analyse information of society about its behaviours, patterns and also similarities and dissimilarities in people using general information.

Text mining tools can help in the collection and analysis of information on the Web. Social scientists typically collect and aggregate such information to discover how society's interests change. New topics that are gaining popularity can be detected from published data on news broadcasters' Web sites. The results from these Web sites are very useful to detect hot topics and gather opinions on such topics.

Now a days shopping on web is very popular. But issues in shopping on the Web are finding the right price and reliability. Prices vary from site to site and it is not practical to visit and scan each site manually. A shopping *agent* can find products and prices given a list of sources and criteria.

Academia is a major producer of information—thousands of research papers, books, and articles are published yearly. Many of this information are available online. To search specific research papers, books, or articles is tedious job. For this an automated crawler can help us. An automated crawler can visit all the Web pages department wise or institute wise or university wise and extract the titles and other relevant information from all faculty publications.

Automated ranking of essays or articles is a program which automatically distinguishes good essays or article from bad ones and assigns grades based on the textual analyses.

In agriculture, many applications using text mining are existing and few more are also possible which helps our farmer to improve quality and production of grains, vegetables, and fruits.

IV. Conclusions

In today's world, we can find many applications of text classification using text mining. Text mining is the automated or sometimes partially automated processing of text. It consists of imposing structure upon text and extracting relevant information or patterns from text. Text classification comprises of identifying common features across documents and organizing those documents into groups based upon the common features. In this paper, various applications of text classification using text mining are discussed which are broadly divided into area like Business, Law, Medicine and Society. Also, architecture of application building is presented.

ACKNOWLEDGMENT

The authors would like to thank all the reviewers for their important and valuable comments that greatly enhance the quality of the paper.

REFERENCES

[1] Aaron M. Cohen and William R. Hersh, *A survey of current work in biomedical text mining*, HENRY STEWART PUBLICATIONS 1467-5463. BRIEFINGS IN BIOINFORMATICS.VOL 6. NO 1. 57–71. MARCH 2005

[2] E.W.T. Ngai, Li Xiu, D.C.K. Chau, Application of data mining techniques in customer relationship management: A literature review and classification, *An International journal of Expert Systems with Applications*36 (2009) 2592–2602

[3] Man Lan, Chew Lim Tan, Jian Su, and Yue Lu, "Supervised and Traditional Term Weighting Methods for Automatic Text Categorization", *IEEE Transactions On Pattern Analysis And Machine Intelligence*, Vol. 31, No. 4, April 2009J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[4] Manisha Pravin Mali , Dr. Mohammad Atique, "A Critical Review of Text Classification using Soft Computing", *ICII 2012 & CSI Annual Convention*, 1-2 Dec 2012, 978-1-25-906170-7

[5] Pang-Ning Tan, Vipin Kumar, Michael Steinbach, *Introduction to Data Mining*, Pearson, 2013, ISBN 978-81-317-1472-0

[6] Thomas W. Miller, *Data and Text Mining A Business Application Approach*, Pearson, 2011, ISBN 978-81-317-2100-1