Chapter 9

Medical Reports Analysis Using Natural Language Processing for Disease Classification

Sumathi S.
St. Joseph’s College of Engineering, India

Indumathi S.
Jerusalem College of Engineering, India

Rajkumar S.
HCL Technologies Ltd., India

ABSTRACT

Text classification in medical domain could result in an easier way of handling large volumes of medical data. They can be segregated depending on the type of diseases, which can be determined by extracting the decisive key texts from the original document. Due to various nuances present in understanding language in general, a requirement of large volumes of text-based data is required for algorithms to learn patterns properly. The problem with existing systems such as MedScape, MedLinePlus, Wrappin, and MedHunt is that they involve human interaction and high time consumption in handling a large volume of data. By employing automation in this proposed field, the large involvement of manpower could be removed which in turn speeds up the process of classification of the medical documents by which the shortage of medical technicians in third world countries are addressed.

INTRODUCTION

Text classification refers to the activity of classifying Natural Language texts from the predefined set of available categories. In a basic way of defining Text classification, it can be defined as the process of finding out generic tags from a document which is a collection of huge number of words. Classification of the document into generic tags enables the users of the application or product to easily access the desired information within the application without facing any kind of problems or troubles. Without the
Medical Reports Analysis Using Natural Language Processing for Disease Classification

usage of these classified tags, it is a tedious activity for the users to navigate through loads of information to locate the required document. Text classification in medical domain could result in an easier way of handling large volumes of medical data. They can be segregated depending on the type of diseases, which can be determined by extracting the decisive key texts from the original document. This classification has the potential to assist both the end-users and the experts involved in the medical domain and hence could improve the efficiency of document classification significantly.

Text classification has been proposed for a long time now and its introduction could ease out things for both users and professionals involved in the application. However due to various nuances present in understanding language in general, a requirement of large volumes of text-based data is required for algorithms to learn patterns properly. So, text classification can be employed whenever there are several classification tags to enable the mapping of large volume of textual data.

The introduction of Text classification in Healthcare could turn out to be productive if they are implemented carefully. One of the major sources for data related to healthcare is the Internet. The number of pages available in the Internet is significantly growing every year which in turn increases the amount of data related to health. Medical information has attracted a wide range of audience since its introduction in the electronic form. Health related information is now widely available for use to the public. They are being provided by several health organizations, medical universities and government institutions in a validated form which can be readily used by the public users.

The users of the medical domain can either be end-users or the professionals involved in medicine. The end-users navigate through the medical documents when they have to search for the desired information related to a particular disease. On the other hand, professional experts require Text classification to ease their hectic processes of locating the documents which may be required during case studies of similar diseases.

One of the popularly used medical database is Medline (Medical Literature Analysis and Retrieval System Online), which constitutes the primary medical repository of the U.S. National Library of Medicine (NLM), including approximately 20 million computer readable records. The above-mentioned data is drastically increasing and is forecasted to gain several more millions in upcoming years. It consists of a rich source of medical and biological information which requires an efficient method of mining to produce useful insights.

Since databases like Medline is used by both layman users and professional experts, classification has to be carried out to satisfy both types of users. The professional experts are well versed with the medical terminologies and tend to use those texts for searching information in the web. But the end-users are generally not aware of the complex medical terms and most often surf the internet using Natural Language terms. The existing systems such as MedScape, MedLinePlus, Wrappin and MedHunt (maintained by HON, the Health on Net Foundation, and a non-profit organization) provide the service of answering queries raised by both classifications of users. But, the problem with the above-mentioned systems is that they involve human interaction.

Initially we require professionals and domain experts to sort the documents manually without which the services cannot run properly. These systems are highly dependent on those experts to carry out the manual classification which has a potential threat of high time consumption. Hence these systems fail when there are large volumes of documents as it becomes impossible to manually classify every single document in a short span of time.