Text Mining: Limitations and Future Scope

Shweta Gupta[#1], Kirti Joshi[#2]
[#]CSE Dept, RIMT-IET, Mandi Gobindgarh, India
[1] shweta5422@yahoo.co.in
[2] kirtijoshi11@gmail.com

Abstract: **This paper explains the concept of text mining. In the first section it explains how text mining is advantageous in different fields in terms of efficiency, development of new knowledge, exploring new horizons, improved research process and many other broader benefits. Further it puts light on barriers, risks and issues involved in text mining. A survey on various issues has been done like legal uncertainties, entry cost, noise, document format, lack of transparency and technical knowledge. The last section discusses the related work that has been done in this path and also provides suggestion to improve the results.**

I. INTRODUCTION

Text mining is a growing new field that attempts to assemble meaningful information from natural language text. It may be loosely characterized as the process of analysing text to extract information that is useful for particular purposes. Compared with the kind of data stored in databases, text is unstructured, amorphous, and difficult to deal with algorithmically. But in modern culture, text is the most common vehicle for the formal exchange of information. Vast amounts of new information and data are generated everyday through economic, academic and social activities, much with significant potential economic and societal value. Techniques such as text and data mining are required to achieve this potential. Businesses use data and text mining to analyse customer and competitor data to improve competitiveness; the pharmaceutical industry mines patents and research articles to improve drug discovery. Economic, academic and social activities generate ever increasing quantities of data. Businesses collect trillions of bytes of information on customer transactions, suppliers, internal operations and indeed competitors; the global research community generates over 1.5 million new scholarly articles per annum; and social networking sites such as Facebook and twitter enable users to share over 1.3 billion pieces of information/content per day. Text mining is required if organisations and individuals are to make sense of these vast information and data resources. The resources need first to be processed i.e. accessed, analysed, annotated and related to existing information and understanding. The processed data can then be mined to identify patterns and extract valuable information and new knowledge. How these information and data resources are analysed depends on their format.

Text mining offers a solution to these problems, drawing on techniques from information retrieval, natural language processing, information extraction and data mining/knowledge discovery as Figure 1 illustrates.

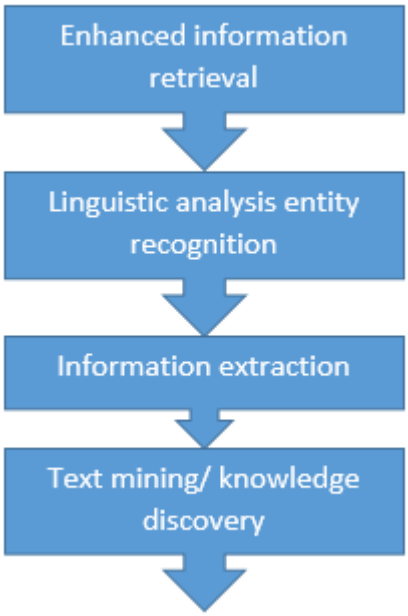

Figure 1 - Overview of the components of text mining

II. VARIOUS ADVANTAGES OF USING TEXT MINING

A. Efficiency

A key benefit of text mining is that it enables much more efficient analysis of existing knowledge. The ability to extract information automatically cuts down the time spent on ensuring coverage of domain knowledge. For example, given the total volume of scholarly publications in the biomedical fields, it could take a human researcher several years to analyse the corpus to identify all relevant sources for a particular problem. Using text mining to identify relevant material could drastically cut down the time required. Further, if the text mined documents were annotated with the semantic information that has been extracted and were then made available for reuse, key resources would be found more quickly.

B. *Unlocking hidden information and developing new knowledge*

The unlocked information can lead to new knowledge and improved understanding. For example, the enormous volume of academic publications means that there may be underlying connections between different subtopics that could not be found without automated analysis. Also the potential links found between diseases and drugs developed for other purposes are a good example of the unlocking of this hidden information.

C. *Exploring new horizons*

In some areas text mining is transforming not just how research is done but also what is researched. New horizons and research questions are being explored. For example, a whole new area of digital humanities has emerged. Research in this area is not only leading to better understanding of the information and social-cultural significance embedded in historical artefacts, it is also providing enhanced tools and methodologies to improve understanding of the multi-media world in which we now live.

D. *Improving research process and quality*

The availability of both text mining tools and the reusable semantic outputs (knowledge representations) is helping to improve the research process itself as they provide new tools and methods that can be applied in innovative ways. For example, a researcher can use text mining to check that their traditional literature review has covered the relevant domain of knowledge. As one researcher reported, this method identified a subset of documents that he had not examined in his traditional literature search. This was because the subset of documents came from a different sub-discipline where they used different terminology for a key concept.

E. *Broader benefits*

The broader economic benefits identified include: cost savings and productivity gains, innovative new service development and new business models. Cost savings and productivity gains from using text mining to explore the scientific research base or consumer data are already in evidence. For example, within the pharmaceutical industry collaborative ventures between traditional competitors explore the existing knowledge base to reduce the costs of drug discovery. New business models may develop for existing business. For example, some copyright holders are exploring allowing their content to be mined for free as hits will attract more visits to their costed service. The potential for societal benefits are significant, particularly with regards to finding drug treatments or cures for serious diseases such as diabetes.

III. DRAWBACKS ASSOCIATED WITH TEXT MINING

This section discusses the various barriers, risks and issues involved in text mining.

A. *Legal uncertainty, orphaned works and attribution requirements*

Permission from the copyright holder is required before the digital copying and annotation required as part of text mining can be undertaken. However, where institutions already have existing contracts to access particular academic publications, it is often unclear whether text mining is a permissible use. The situation is further complicated where there are orphaned works, where the rights holder is unknown or cannot be contacted. Further, as the law currently stands, copyright law can be overwritten by contract law. So even if, there were to be an exception that allows text mining of copyrighted materials for non-commercial research, there could still be considerable restrictions, leading to uncertainty. Even where text mining is allowed within publisher contracts, licensing terms that require the full attribution of derivative works developed in the text mining process can effectively prevent text mining usage.

B. *Entry costs*

The entry costs associated with development and training of text mining tools for use within a different topic from that for which they were originally designed were also identified as a significant barrier to uptake of text mining. Investment in training for researchers is also required.

C. *Noise in text mining results*

Text mining of documents may produce errors. False connections may be identified or others missed. In most contexts, where the noise or error rate is sufficiently low, the advantages of automation outweigh the possibility of a higher error than that produced by a human reader.

D. *Document formats*

The format of many documents also limits the amount of text that can be mined. This is particularly an issue when the documents are stored as images or 'pdfs', as it is difficult to identify and extract relevant metadata. There is no standard fully automated way to convert such documents into more text mining-friendly formats. Further, where these more friendly formats are available, publishers may impose additional charges for access. The tendency to store papers lodged within institutional repositories as pdfs only further contributes to the problem. XML is the preferred format for text mining.

E. Lack of transparency

For many, text mining is observed as a black box where corpora of text documents are input and new knowledge is output. Where researchers do not have the technical knowledge or skills to understand the internal workings of text mining, or do not have access to the corpora or text mining tools, text mining is effectively opaque. This lack of transparency limits use in three ways. First, it discourages researchers from using what they do not fully understand. Second, without good understanding of the process involved, the potential of new and innovative applications may be missed. Third, if the process and research data are not transparent then it is impossible for others to reproduce the results.

F. Lack of support, infrastructure and technical knowledge

Text mining is a highly specialised activity, which creates additional annotated copies of corpora and large information repositories. For small research groups or individual researchers, lack of a central infrastructure to support this may rule out use of text mining. Consultees in non-scientific areas also felt that it was difficult to obtain funding for technical infrastructure and support.

IV. RELATED WORK

There exists many techniques for text mining like clustering, classification, association and so on. Selection of the technique depends on the area on which it needs to apply and there are many algorithms to implement each technique. Somewhere more than one technique is applied to refine the results. S. Subbaiah [1] has combined clustering and classification to overcome the problem of false indexing. Also most of the algorithms suffer with overlap in identifying the category of the document.

V. FUTURE SCOPE

Although the current techniques has improved the previous results but still there is a vast scope to improve the accuracy of the results of text mining. The problem of overlapping can be reduced if two or more different algorithms are either clubbed together or used one after the other to overcome (or to reduce) the shortcomings of text mining.

I. References

[1] Subbaiah , S. "Extracting Knowledge using Probabilistic Classifier for Text Mining " Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, February 21-22.

[2] Selvakumar, A. "An Adaptive Partitional Clustering Method for Categorical Attribute using K-Medoid" IJCSMC, Vol. 2, Issue. 4, April 2013, pg.197 – 204

[3] http://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining

[4] Kaur, Supreet "A Survey on Various Clustering Techniques with K-means Clustering Algorithm in Detail" IJCSMC, Vol. 2, Issue. 4, April 2013, pg.155 – 159

[5] http://www.cs.waikato.ac.nz/~ihw/papers/04-IHW-Textmining.pdf

