International Journal of Computer (IJC)
ISSN 2307-4523 (Print & Online)

http://ijcjournal.org/

A Review of Text Mining Techniques and Applications

Kanak Sharma[a]*, Ashish Sharma[b], Dhananjay Joshi[c], Nikhil Vyas[d], Arpit Bapna[e]

[a,b,d,e]*B.Tech.(CE), NMIMS University, Maharashtra 425405, India.*
[c]*Asst. Prof., CE Department, NMIMS University, Maharashtra 425405, India.*
[a]*Email: sharmakanak33@gmail.com*

Abstract

Due to the ever increasing rate at which information is generated, text mining and its automated analysis have become the need of the hour. The paper discusses some of the developments in text mining applications, primarily reviewing techniques in the classification, summarization and analysis of text, as advocated by academia. The goal is, in essence, to ultimately turn unstructured text into useful data and information for analysis using critical methods. We introduce the paper by introducing the concept of "textual analysis" similar to text mining done using the analysis of Natural Language texts, their respective techniques in use and the open source tools in use to do so. We survey varied topics that use NLP, and also expand the horizons of this domain by devising new techniques for improving the efficiency even in limited amounts of data, improved accuracy, new methods, novel approaches, and new application areas for it, and relating to text summarization and text classification. Various text mining techniques used in text classification and summarization are reviewed, followed by the application areas of text mining being worked upon by businesses. Finally, the paper concludes by introducing "organizational text mining" and emphasizing the need for it.

Keywords: natural language processing; text mining; text classification; text summarization.

1. Introduction

NLP is defined as a domain of CS in which the algorithms and techniques are used to comprehend and create natural language. Like every other limb of computer science, NLP has also been paired with Machine Learning (ML), to automate classification and pattern discovery in electronic documents and other unstructured text mined from various sources, probably for the best.

---

* Corresponding author.

Advanced algorithms, neat NLP techniques and humongous amounts of data have ensured that Natural Language processing as a field of study is progressing rapidly, now more than ever. Text mining is the process of extracting text from various sources, converting it into structured information, determining relationships between them and subsequent analysis of those relationships between the lemmas for finding patterns, solving problems and creating useful applications of the same. "Textual Analysis" can be formulated as an umbrella under which evaluation, use and applications of Natural language Processing and computational linguistics lie. It encompasses text extraction [11], pre-processing, classification, summarization and a lot of other activities that can be performed with NLP. Text mining is being done by various algorithms ranging from the widely cited Naïve Bayes to the relatively unknown techniques of back propagation in Artificial Neural Networks (ANN).

2. Text Mining

Text mining is the process of extracting invaluable information from text [15]. Text mining study is constantly gaining more and more reputation recently because of accessibility of the growing variety of sources and count of electronic documents. The resources of semi-structured and unstructured information in the world include the WWW, news articles, biological records, governmental electronic repositories, online forums, digital libraries, chat rooms, and electronic mail and blog repositories. Hence we can say that proper knowledge discovery from these resources is a research area of some importance. The paper on Document classification methodology [1] establishes the idea of attaining high accuracy in classifying documents. It focuses mainly on small training data. It accomplishes this by compiling results from previous work under large data sets that use Bayesian classification and statistical decision theory. The first half uses large data set and estimates the unknown class x in the new document under the condition that the string of key words y` n` in the doc and the learning data named doc `L are given where a loss function gives a binary output when decision function gives an estimate i.e. 1 or 0. In the next set, it works on small training data and documents occurring from different sources as a means of estimating data using Dirichletian distribution. For the new classification method, accuracy is higher than before when the quantity of documents in the training data was small, and is almost the same when the training data is big, but parameters A1, A2 and A3 for the working of this method have to be taken heuristically.

Another research paper on Sentiment expression via emoticons [2], in which Hao Wang argues the idea of emoticons as strong signal of sentiment on social media and the clarity they bring to sentiment polarity and expression. First, to display the occurrence of emoticons, it discovers that out of 1.5 billion tweets, exactly 8,625,753 of emoticons were found. Four analyses are done to examine relationship between emoticons and the backgrounds wherein they are used. The 1st study graphed recurrent ones. The next study inspected clusters of words and the meaning conveyed by the emoticons. 3rd analyzed the emotion spread of texts before and after smileys were deleted from text. The 4th one, showed the theory that deleting smileys in text affects emotion arranging. The results established that only a few smileys are tough sentiment signals, and that a large group of emoticons convey difficult sentiments therefore should be treated with caution. A paper by Shweta and Sonal Patil [3] describes techniques for automatic marking of free-text responses using Natural Language Processing they mention them as being prorated into three main categorizes: Firstly, the straight forward Statistical Technique based on keyword matching. It lacks the ability to tackle problems on various fronts such as

synonyms, accounting for the order of words, or dealing with lexical variability. Information Extraction (IE) Technique on the other hand, consists of getting structured information from free text in order to extract dependencies between concepts breaking the text into concepts and their relationships and then comparing dependencies against human experts to reach the decision. Finally, the Full Natural Language Processing technique involves parsing of text and finding semantic meaning of text and finally comparing it with subject text and assigning final scores. In Design of an Automated Essay Grading (AEG) system in the Indian Context [4], the researchers deal with the problem of scoring systems also known as Automated essay grading systems to automatically assess the answers of students in exams like TOEFL where students write essays which are presently being assessed by both – AEG and a human and an average is taken. Dealing with linguistics has an inherent problem that is really complex to deal with, multilingual contextual recognition. Currently, the systems used for grading are essay grade systems which deal with pure English essays or ones printed in pure European languages. We have 21 regional languages and pressure of these local languages, in English, is highly observed. Newspapers in Hyderabad, India sometimes print– "Now the time has come to say 'albida' (good bye) to monsoon" [4]. Due to influence of regional languages such as Hindi or Bangla on non-native English speakers the consequence of TOEFL exams, has revealed lower scores alongside Indian students and other Asian students as far as the Essay section is concerned. A review paper on text summarization defines transcript summarization as a course of extracting or collecting significant information from unique text and presenting that information in the shape of a synopsis. This paper is an effort to present the sight of text summarization from every facet in a historical review paper format [5]. The method arranged for summarization varies from structured, for all time being used to begin with, being the simplest to comprehend and intuitive to linguistic. Further it brings to light that in India, multi linguistic techniques are being explored and work has been done, but presently it is in an infancy state. This paper gives a theoretical sight of the present situation of study for transcript summarization [9-10]. Another beautiful and informative paper on featured-based sentiment classification for hotel reviews using Bayesian classification talks about facts and opinions and how sentiment analysis and text mining on the data on the Internet, specifically in hotel reviews, can be used to classify positive and negative reviews [6]. The paper talks in detail about using the following techniques to achieve the automated classification objective [8] such as firstly, semantic orientation or synonym- based review classification, secondly, ML-based classification using techniques such as kNN, SVM and Naïve Bayes, followed by a third approach of using using NLP techniques such as NER as POS tagging and finally JAPE rule which is essentially a set of pattern action rules.

2.1. Text Mining Techniques

Text Summarization and classification being the most popular applications of text mining, especially amongst businesses, it is only fair that we discuss some of the techniques used to implement them.

2.1.1. Text Summarization

Various research papers talk about text summarization and its urgent need for business process automations and intelligent systems implementations [9-10]. The main techniques of text summarization are abstractive and extractive text summarization. Abstractive summarization generates summaries that are normally broadly classified into two groupings, structure based approach and semantic based approach.

- Structured Based Approach: Structured based approach codes most vital substance or information from the text during cognitive schemes such as patterns, extraction policy and other arrangements such as hierarchy, ontology, and guide and body phrase makeup.
- Semantic Based Approach: In Semantic support approach on the other hand, semantic depiction of text is used as input into natural language generation (NLG) system. This system focus on recognizing the variety of noun and verb phrases by dispensation of linguistic data.

The extractive summarization way consists of choosing significant sentences, paragraphs etc. from the unique text and concatenating them into smaller form. The significance of sentences is determined based on arithmetic and linguistic features of sentences, frequently on the basis of priority assigned.

2.1.2. Text Classification

The different type of classification models are decision trees, neural networks (NN) and genetic algorithm (GA) [7]. Classification using Decision Trees can be done by three major techniques:

- C4.5 Algorithm: Generating a classification decision tree for the given data set by recursively partitioning of the given input data. The algorithm considers all the possible tests that can split the data set and selects a test that gives the best information gain. It should be noted that the decision tree is grown using DFS strategy.
- Sequential Decision Tree based Classification: A decision-tree model consisting of internal decision leaves and nodes. It consists of tree induction and pruning. The inferred decision-tree is made further robust and concise by removing all statistical dependencies on the original training data set.
- Synchronous Tree Construction Approach: using multi-processor architecture for fast and efficient decision tree construction and expansion.

In classification using neural network, interlinked processing nodes are used for doing the classification. An artificial neural network is a mathematical model inspired by biological neurons consisting of an interconnected group of artificial neurons processing information using a connection-oriented approach to computation. We are given a set of sample pairs and the aim is to find a function that matches the sample, that is, we wish to infer the mapping implied by the data. A cost function is used, which is related to our mapping and the data and it implicitly contains prior knowledge about the problem domain. The function must be such that it correctly buckets or classifies all input data up to a certain degree of error. Neural networks can exist in two major forms:

- Multi-Layer Perceptron is a form of simple feed forward NN, also referred to as a MLP. The neurons are stacked layer-wise with outputs always flowing toward the output layer. If only one layer exists, it is called a perceptron.
- Back Propagation algorithm is a technique that accommodates weights in neural network by making weight changes backwards from the output to the input nodes.

Classification using genetic algorithms works on the same basis as biological evolution works in species as

mentioned in theories on evolution. In Genetic Algorithm, the units or individuals are called chromosomes. After the initial population is generated randomly, selection and permutation function is run so that the termination criterion is finally reached. The selection operator is intended to improve the average "quality" of the population by giving individuals of higher quality, a higher probability to be passed onto the next round of selection and mating, as is the case with humans. Each execution, i.e. one loop is called a generation. The quality of an individual is measured by a fitness function. Genetic Operators namely, mutation and crossover are applied to generate offspring from the existing population. Genetic Algorithms need a termination criterion to stop the complete process. If no "significant" improvement is observed, in consecutive generations, the entire process is stopped. The sufficiency criterion is decided by developer according to the problem domain.

2.2 Text Mining Applications

Text mining is now used in a wide array of research, business and government needs. Applications can be sorted into a variety of categories by business function or analysis type. Classifying solutions in this manner, the numerous currently used application categories include:

- Enterprise Business Intelligence: analyzing data to predict market trends and to improve enterprise performance [12].
- Sentiment Analysis: It generally refers to the use of text analysis, natural language processing and computational linguistics to find and extract subjective information from various sources regarding behavioral sciences [2].
- Natural Language/Semantic Toolkit or Service: huge open source libraries such as the Stanford NLTK [14] toolkit and Apache openNLP library provide the functionalities of NLP in a comprehensive package ranging from features such as Part-of-Speech (POS) Tagging to Named Entity Recognition (NER) to Lemmatization to Chunking and so on.
- Social media monitoring: Term frequency-inverse document frequency (TF-IDF) analysis and relative normalized term frequency analysis are pretty common to identify the trending topics in social media web sites such as Twitter and Facebook [13].

3. Conclusion

After reading, summarizing, categorizing and contemplating upon research work done in the field of Natural Language Processing with primary focus on textual mining/ analysis we feel that the majority of research being done is consumer focused. The textual data being collected is produced by consumers either on social media by users, or by test-takers, in case of examinations. Hence, there should be more research focused on using text mining and NLP techniques to analyze the hosts of this textual data, like Facebook or Twitter or TOEFL. Various social media alerts are automatically generated by recommendation algorithms these companies employ. And as far as examination question answer grading is concerned, an analysis of the question paper itself could be done, this can be used to find out how efficient a system is, such as the Friends' posts' recommendation notifications given by Facebook could be improved via sentiment analysis of users posts or a complete Bloom's Taxonomical evaluation of major competitive exams like SAT, TOEFL, IELTS and GRE can

be done and the efficiency of them mapping to the careers of their respective test takers can be calculated. Hence, what we would coin as “organizational text mining” is currently the need of the hour, whether it is related to data produced by large social media corporations or other large organizations and bodies, what it says about them as opposed to what it does about their users, needs to be focused on and researched with greater interest and more attention.

References

[1] Yasunari Maeda, Hideki Yoshida, and Toshiyasu Matsushima. “Document classification method with small training data,” in Proc. ICCAS-SICE, 2009.

[2] Hao Wang and Jorge A. Castanon. “Sentiment Expression via Emoticons on Social Media” in Proc. IEEE International Conference on Big Data, 2015.

[3] Shweta Patil and Sonal Patil. "Intelligent Tutoring System for Evaluating Student Performance in Descriptive Answers Using Natural Language Processing." International Journal of Science and Research, 2014.

[4] Siddhartha Ghosh and Dr. Sameen S Fatima. “Design of an Automated Essay Grading (AEG) system in Indian Context.” International Journal of Computer Application, vol.1, No.11, 2010.

[5] Deepali K. Gaikwad and C. Namrata Mahender. “A Review Paper on Text Summarization”. International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Issue 3, Mar. 2016.

[6] Tushar Ghorpade and Lata Ragha. “Featured Based Sentiment Classification for Hotel Reviews using NLP and Bayesian Classification” presented at the International Conference on Communication, Information & Computing Technology (ICCICT), Mumbai, India, Oct. 2012.

[7] Bhumika, Prof Sukhjit Singh Sehra and Prof Anand Nayyar. “A Review Paper On Algorithms Used For Text Classification”. International Journal of Application or Innovation in Engineering & Management, Vol. 2, Issue 3, March 2013.

[8] Mita K. Dalal and Mukesh A. Zaveri. “Automatic Text Classification: A Technical Review”, 2011.

[9] N. Moratanch and Dr. S. Chitrakala. “A Survey on Abstractive Text Summarization”, in Proc. International Conference on Circuit, Power and Computing Technologies, 2016.

[10] Urmila Shrawankar and Kranti Wankhede. “Construction of News Headline from Detailed News Article”, 2016.

[11] Manju Khari, Amita Jain, Sonakshi Vij and Manoj Kumar. “Analysis of Various Information Retrieval

Models", 2016.

[12] B. Azvine, Z. Cui, D.D. Nauck and B. Majeed. "Real Time Business Intelligence for the Adaptive Enterprise", 2006.

[13] James Benhardus. "Streaming Trend Detection in Twitter", 2013. Benhardus, James, and Jugal Kalita. "Streaming trend detection in twitter." International Journal of Web Based Communities, pp. 122-139, 2013.

[14] Steven Bird. "NLTK: The Natural Language Toolkit", Proc. COLING/ACL on Interactive presentation sessions, pp. 69-72, 2006.

[15] Chetan Botre, Saad Patel, Shrinivas Kunjir and Swapnil Shinde. "NoteMate - A Note Making System Using OCR and Text Mining" in International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 3, Mar. 2015.