Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.778199

AN INNOVATIVE TEXT CLASSIFICATION BASED ON MACHINE LEARNING TECHNIQUES

Dr.S.Rizwana
Assistant Professor and Head, Department of Computer Science(SF),Erode Arts and Science College, Erode

Mrs.N.Sasikala
Assistant Professor, Department of Computer Science (SF), Erode Arts and Science College, Erode

Abstract

Text mining is a growing new field that attempts to glean meaningful information from natural language text. It may be roughly characterized as the process of analysing text to extract information that is useful for particular purposes. Most of soft documents are stored as text, text mining is believed to have a high viable potential value. Knowledge may be exposed from many sources of information yet, unstructured texts remain the largest readily available source of knowledge .Text classification classifies the documents according to predefined categories .This research work gives the introduction of text classification, process and techniques of text classification as well as the overview of the classifiers and tried to compare the some existing classifier on basis of few criteria like time complexity, principle and performance.

Keywords: Feature Extraction, Future Selection, Machine Learning (ML), Natural Language Processing (NLP), Structured Date, Text Mining, Unstructured Date.

1. INTRODUCTION

Nowadays people yield an extremely huge measure of data daily. Everyone make bulky data's from wake up morning to bed at night. Each one can get and produce data's from phone ,email's, messages, start scrolling through some of various applications .All this is unstructured, raw data. Without automated text mining, it is useless, because it's unsearchable, there are no shapes identified, no keywords grew. Structured data is highly systematized and easily understood by machine language. Example Name, Age, Gender, Address, Phone number, Currency, Date, Billing information, etc. Those working within relational databases can quickly input, search, and manipulate structured data using a relational database management system (RDBMS). This is the most attractive feature of structured data. But once user start text mining and organising this data, suddenly it becomes something useful for our needs.at the same time in our network world unstructured datas are moving like SQL database, Spread sheets, Sensor data's ,inputs from medical devices, online forms,etc.The following figure shows the percentage of data's from various resources in the universe.

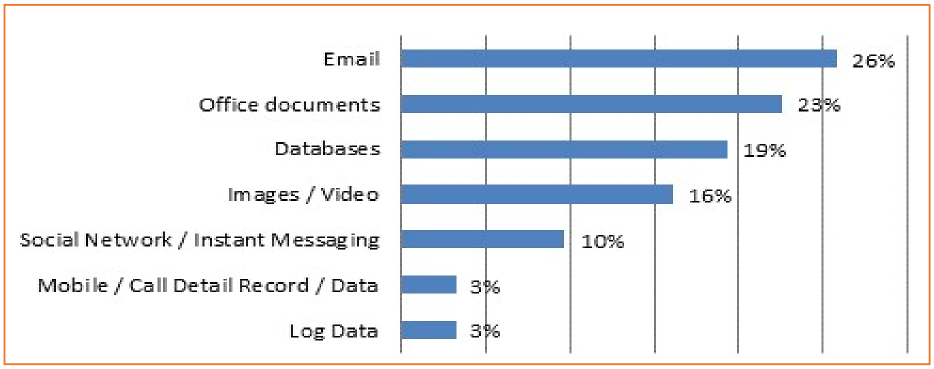

Fig 1 : Various Data Sources and its range

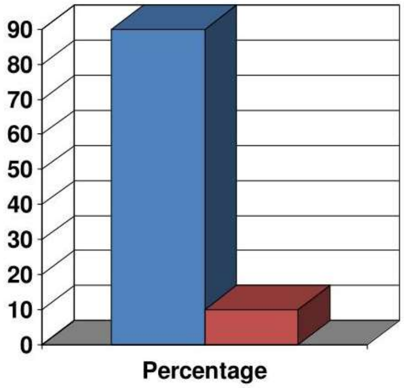

Fig 2: Data vs. Structured Data

Table 1: Sources of Data and its Type.

SNO	**Data Types**			
	Structured Data		Unstructured Data	
	10%		90%	
	Sources of Data	Percentage of Data (%)	Sources of Data	Percentage of Data (%)
1	Web and Server Logs	7	Email	26
2	Spread Sheets	4	Documents(Office)	23
3	Addresses	3	Databases	19
4	Sensors	8	Images & Videos	16
5	Medical Devices	6	Social Network	10
6	Stock Information	2	Mobile	3
7	Geolocation	1	Log Data	3

The Figure 2 shows the standard data from total data over the network. The text mining studies are gaining more importance recently because of the availability of the increasing number of the electronic documents from a variety of sources. Which include unstructured and semi structured information. The main goal of text mining is to enable users to extract information from textual resources and deals with the operations like, retrieval, classification (supervised, unsupervised and semi supervised) and summarization Natural Language Processing (NLP), Data Mining, and Machine Learning techniques work together to automatically classify and discover patterns from the different types of the documents. The above Table 1 displays types of data and its sources.

II. Literature Review

Data is a special one, it is a group of facts and details like numbers, text, figures, symbols and annotations, that does not contains any specific meanings. In simple terms, the user can conclude that data is an unorganised description of raw facts from which information can be extracted [3]. Text mining is a process of eradicating interesting and non-trivial patterns from huge amount of text documents. The field of text mining usually deals with texts whose function is the communication of factual information or opinions, and the motivation for trying to extract information from such text automatically is compelling—even if success is only partial[7].

Text mining is a process of extracting interesting and non-trivial patterns from huge amount of text documents. The information age is characterized by the fast growth of data, mostly unstructured data. Unstructured data is often text-heavy; including news articles, social media posts, Twitter feeds, transcribed data from videos, as well as formal documents. The availability of this data presents new chances, as well as new challenges, both to researchers and research institutions. In this paper, user review several existing methodologies for analysing texts and introduce a formal process of applying text mining techniques using the open-source software. In addition, User discuss to possible empirical applications. Nowadays, research in text mining has become one of the wide spread fields in analysing natural language documents Nowadays, research in text mining has become one of the widespread.

fields in analyzing natural language documents. Text mining is about natural language textual data that is stored in semi designed and unstructured data format. Text mining processes are used interruptedly in business, researchers and the web applications, on the internet and many other areas [1], Research in Text Mining has become one of the widespread fields in analysing natural language documents. The procedure of text mining begins with gathering documents through different resources [9]. Generally, a classification technique could be divided into statistical and machine learning (ML) approaches. Statistical techniques purely satisfy the proclaimed hypotheses manually, therefore the need for algorithms is little, but ML techniques were specially invented for automation [6]. In recent years, there has been a lot of progress in natural language modelling and representation. NLP is of major interest in research as it represents the core business of Internet companies today [7]. This paper offers a primer on how to systematically extract quantitative information from unstructured or semi-structured text data. Quantitative representation of text has been widely used in disciplines such as computational phonology, sociology, communication, political science, and information security. However, there is a growing body of fiction in economics that uses this approach to analyze macroeconomic issues, particularly central bank communication and financial stability. The use of this type of text analysis is growing in popularity and has become more widespread with the development of technical tools and packages facilitating information retrieval and analysis.Nowadays, research in text mining has become one of the widespread. There are different techniques and methods for TM in order to find new structures, patterns, or associations. Some TM has involved the assumption of an a priori classification (preprocessing) into attributes and then proceeded via “classical” DM methods, i.e. statistical analysis, associations, etc [5], [2]. Others, investigate the full text of document collection, e.g. categorization used above, or purely analytical results. A common end-goal of much TM is a more efficient, complete, and/or specific way to browse and search large collections of documents. Thus, the main techniques in TM can be divided according to the tasks they perform in the discovery process: the kind of information they extract and the kind of analysis/association done with them. The following table notifies methods and techniques are used in this research for produce quality results in every process of text mining.

III. **Methods and Techniques**

Table 2: Text Mining and Its Techniques

SNO	Text Mining Processes	Methods &Techniques
1	Information Extraction	Intelligence Techniques-Filters(Media & Weighted Median Filters)Natural Language Processing(NLP)Functions(Left,Right,Mid)
2	Information Retrieval	Pattern recognition ,Analytical processes
3	Natural Language Processing	Co-referencing Technique
3	Gathering & Clustering	Fuzzy Means
4	Text Summarization	Heuristics methods
5	Text Classification	Machine Learning Algorithms and Knowledge Engineering Techniques

A. Information Extraction

Information Extraction (IE) is a technique that quotations expressive information from large amount of text. Domain experts specify the attributes and relation according to the domain . IE systems are used to extract specific attributes and entities from the document and establish their association [8]. The extracted corpus is stored into database for additional processing. Precision and recall process is used to check and evaluate the relevance of results on the extracted data. In-depth and complete information about the relevant field is required to perform information extraction process to attain more relevant results [9]. There are two terms used in making or extracting some relevant information from a data-set, i.e., prediction modelling, and text mining [7]. The following extraction and section processes are the important for extracting meaningful words from documents.

- Feature Extraction – In this process, we try to develop some new features from existing ones. This objective can be achieved by analysing an existing feature or combining two or more features based on some mathematical operation.
- Feature Selection – In this process, we try to reduce the dimensionality of the dataset which is generally a common issue while dealing with the text data by selecting a subsection of features from the whole dataset.

B. Information Retrieval Information Retrieval (IR)

IR is a process of mining pertinent and associated patterns according to a given set of words or phrases. There is a close relationship in text mining and information retrieval for textual data. In IR systems, different algorithms are used to track the user's performance and search relevant data accordingly [9]. Google and Yahoo search engines are using information retrieval system more frequently to extract relevant documents according to a phrase on Web. These search engines use inquiry based algorithms to track the trends and attain more significant results. These search engines provide user more relevant and appropriate information that satisfy them according to their needs [8].

In the process of Information recovery, user try to process the available documents and the text data into a structured form so, that user can apply different pattern recognition and analytical processes[11]. It is a process of extracting relevant and associated designs according to a given set of words or text documents. For this, user has processes like Tokenization of the document or the curtailing process in which user try to extract the base word or let's say the root word present there.

C. Natural language processing (NLP)

NLP concerns to the automatic processing and analysis of unstructured textual information. It performs different types of study such as Named Entity Credit (NER) for abbreviation and their synonyms extraction to find the relationships among them [10]. NER identify all the instances of specified object from a group of documents. These entities and their cases let the identification of relationship and other information to attain their key concept. However, this technique lacks whole dictionary list for all named entities used for identification [3], [2]. Complex query based algorithms need to be used to attain acceptable results. In real world, a single entity has many terms like TV and Television. Sometimes, a group of following words have multi-word names to identify the borders and resolve meeting issues by using classification technique. Approaches to deal with NER usually fall into four categories: lexicon, rule, statistical based or mixture of these approached. NER systems have achieved the relevance level from 75 to 85 per cent [2]. To extract synonym and abbreviation from textual data, co-referencing technique is frequently in use for NLP. Natural Languages (NL) have lot of difficulties as a text extracted from different sources doesn't have identical words or contraction. There is a need to detect such issues and make rules for their uniform identification [3]. For example, NER and co-referencing approaches establish a logical relationship to extract and identify the role of person in an organization (use the name of a person at once and then use pronoun instead of name again and again) [2].

D. Gathering

Gathering is an unsupervised process to classify the text documents in groups by applying different bunching algorithms. In a group, similar terms or designs are grouped extracted from various documents. Clustering is performed in top-down and bottom up manner. In NLP, various types of mining tools and techniques are applied for the analysis on unstructured text. Different techniques of gathering are hierarchical, supply, thickness, centroid, and k-mean [2].

E. Text Summarization

Text summarization is a process of collecting and producing concise representation of original text documents . Pre-processing and processing operations are performed on the raw text for summarization. Tokenization, stop word removal, and stemming methods are applied for pre-processing. Lexicon lists are generated at processing stage of text summarization. In past, automatic text summarization was performed on the basis of occurrence a certain word or phrase in document. Later on, additional methods of text mining were introduced with standard text mining process to improve the relevance and accuracy of results. To summarize the text documents, weighted heuristics method extract features by following specific rules. Sentence length, fixed phrase, paragraph, thematic word, and upper case word identification features can be implemented and analysed for text summarization. Text summarization techniques can be applied on multiple documents at the same time. Quality and type of classifiers depend on nature and theme of the text documents [11**].**

F. **Text classification**

Text classification (TC) is an important part of text mining, looked to be that of manually building automatic TC systems by means of knowledge-engineering techniques, Manually defining a set of logical rules that convert expert knowledge on how to classify documents under the given set of categories[10]. For example would be to automatically label each incoming news story with a topic like “sports”, “politics”, or “art”. a data mining classification task starts with a training set D = (d_1 ….. d_n) of documents that are already labelled with a class C1,C2 (e.g. sport, politics). The task is then to determine a classification model which is able to assign the correct class to a new document d of the domain Text classification has two flavours as single label and multi-label .single label document is belongs to only one class and multi label document may be belong to more than one classes In this paper we are consider only single label document classification.

Conclusion

The accessibility of massive volume of text centred data want to be observed to excerpt treasured information. Text mining techniques are recycled to analysed the stimulating and applicable information successfully and efficiently from great volume of unstructured data. This paper presents a brief overview of text mining techniques that help to improve the text mining process. Specific outlines and orders are practically in order to extract valuable information by eradicating immaterial details for prognostic analysis. Assortment and use of accurate techniques and tools giving to the province help to make the text mining process easy and efficient. Domain knowledge incorporation, fluctuating concepts granularity, trilingual text refinement, and natural language processing ambiguity are most important issues and galas that arise during text mining process. In future research work, we will focus to design algorithms which will help to resolve issues presented in this work.

References

1 Akshaya Udgave , Prasanna Kulkarni,” TEXT MINING AND TEXT ANALYTICS OF RESEARCH ARTICLES”,PalArch’s Journal of Archaeology of Egypt/Egyptology”, 17(6). ISSN 1567-214,2020.

2 Ian H. Witten,”Text Mining”, Computer Science, University of Waikato, Hamilton,

New Zealand.

3 Jose Joaquin Mesa-Jiménez „ Lee Stokes, Qingping Yang , Valerie N. livina ” MACHINE LEARNING FOR TEXT CLASSIFICATION IN BUILDING MANAGEMENT SYSTEMS”, Journal of Civil Engineering and Management, Volume 28 ,Issue 5, 408–421, 2022.

4 Jonathan Benchimol a, Sophia Kazinnik b , Yossi Saadon a,”Text mining methodologies with R: An application to central bank texts”,Machine Learning with Applications Volume 8, 15 June 2022.

5 Said A. Salloum, Mostafa AI- Emran,Azza Abdel Monem ,Khaled Shaalan,”Using Text Mining Techniques for Extracting Information from Research Articles”, Intelligent Natural Language Processing: Trends and Applications,pp.373-397,2018.

6 Shabnam Kumari, V. Vani, Shaveta Malik, Amit Kumar Tyagi,Sravanti Reddy,” Analysis of Text Mining Tools in Disease Prediction”, International Conference on Hybrid Intelligent Systems, **Hybrid Intelligent Systems** ,pp 546–564,2020.

7 Ramzan Talib, Muhammad Kashif Hanif, Shaeela Ayesha, Fakeeha Fatima,” Text Mining: Techniques, Applications and Issues”, International Journal of Advanced Computer Science and Applications, Vol. 7, No. 11, 2016

8 Thangaraj M,Sivakami M,” TEXT CLASSIFICATION TECHNIQUES: A LITERATURE REVIEW”,Interdisciplinary Journal of Information,Knowledge and Management,Vol 13,2018.

9 Vandana Korde,C Namrata Mahender” Text Classification and classifiers: A Survey”, International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012.

10 Vishal Gupta, Gurpreet S. Lehal,” A Survey of Text Mining Techniques and Applications”, Journal of Emerging Technologies In Web Intelligence, Vol. 1, No. 1, August 2009.

11 1. YU ZHANG, MENGDONG CHEN,LIANZHONG LIU,” A REVIEW ON TEXT MINING”, IEEE INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING AND SERVICE SCIENCE (ICSESS), 2015.