ISSN 2063-5346

ANALYSIS OF TOOLS AND ORGANIZATION MINING RESEARCH PAPER

Akshat Aman[1], Suman Anand[2], Priyansh Brannen Lall[3], Aakansha Arora[4], Himanshi Metta[5], Pawan Kumar Patnaik[6]

Article History: Received: 10.05.2023 Revised: 29.05.2023 Accepted: 09.06.2023

Abstract

As the trend increases of publishing and storing research papers in various platforms available online to read, to take reference and further carry on the research, the mining of research papers plays an important role in today's era. There are multiple tools and organizations feeding the above purpose. These tools help in mapping and connecting the required data and fetching the useful information present in a large number of research papers published in the same category. Despite the ambiguity of textual data, the organizations are helping researchers by using text mining tools to answer the research questions and find the solution to emerging problems and technologies. This paper aims in listing and comparing the available tools in the public domain and the organization promoting this aim and cause. As the trend continues there have been several research questions arising like, the tools available for the process, their limitations, the organizations that bolster these tools and technology, domain-specific challenge, its accuracy, etc. In this paper, we have tried to answer some of the questions listed above.

Keywords: *Text Mining, Information Retrieval, Natural Language Processing, Scientific Paper Mining, Machine Learning, Research Questions.*

[1,2,3,4,5]Research Scholar, Dept. of Computer Science & Engineering, Bhilai Institute of Technology, Durg, C.G, India
[6]Associate Professor, Dept. of Computer Science & Engineering, Bhilai Institute of Technology, Durg, C.G, India

[1]aman.akshat07@gmail.com, [2]sumananand222@gmail.com, [3]priyansh.blessed@gmail.com, [4]aakansha.a03@gmail.com, [5]himanshimetta@gmail.com, [6]pawanpatnaik37@gmail.com

DOI:10.48047/ecb/2023.12.9.08

1. Introduction

Text Mining is a common process in which we extract the information required using a set of

Documents[2]. It provides basic preprocessing methods, such as identification, extraction of representative characteristics, and advanced operations such as identifying complex patterns [2][3]. Document classification is a task that consists of assigning a text to one or more categories:

the main topics and name of its class of subject.

Recently, the ever-growing availability of datasets and research papers in machine-readable formats and present online freely published by authentic publishers has made possible a change in perspective in the field of bibliometrics and text mining. From movements such as Open Science and preprint databases to the Open Access, the development and emerging of online platforms such as ArXiv, CiteSeer or PLoS, and so forth, largely contribute to facilitating the experimentation with datasets of articles, making it possible to perform text and paper mapping and finding out useful information from bundles of a research paper.

The field of mining research paper offers answers to valuable and undiscovered research questions. Recently applied open-source tools for text processing for such tasks include NLTK, Mallet, OpenNLP, CoreNLP, Gate, CiteSpace, AllenNLP, and others. e.gPubMed OA, CiteSeerX, JSTOR, ISTEX, Microsoft Academic Graph, ACL anthology provide data to these communities for free of cost. Knowledge resources alone are not capable of capturing different language variations. Text mining technologies combine knowledge resources, linguistic analysis, and machine learning to deal with the problem. Furthermore, text mining tools can extract terms from text, and also the relationships between the information and data.

The motivation behind our review paper is to analyze various tools/software present in the public domain which helps in mining scholastic articles and papers and help in searching, classifying, and recommending required scientific literature automatically. In this paper, we have tried to answer some of the most common questions which people wonder about the topic.

- What are the latest and most popular tools/software in use by practitioners for mining scientific publications?
- What are their limitations?
- How are these tools helpful to the researchers and which tools can be used in what scenarios?

The process of manually extracting and reviewing the most relevant literature according to our research requirements needs manual effort and is time-consuming. The scope for error also increases while searching manually, and thus we need tools or software which does this tedious job for us in real-time.

2. Working of Text Mining: Overview

The steps involved in the process of text mining are mentioned below:

2.1 Text preprocessing

The text preprocessing step is further categorized into:

a.Tokenization

The process of parsing sensitive data into non-sensitive data known as "tokens", that can be used in a database or internal system without bringing it into scope is termed tokenization. It can be used to secure sensitive data by replacing the

original data with unrelated information of the same length and format type. This step segments the whole text into words by removing blank spaces, commas, etc.

b.Stopword Removal

This step involves removing HTML, XML tags from web pages. Then the process of removal of stop words such as 'the', 'are', 'of' etc is performed.

c.Stemming

Over-stemming occurs when a much larger part of a word is chopped off than what is needed, which removes the redundant word, reducing it to the same root word or stem incorrectly while they should have been reduced to two or more than two stem words.

2.2 Text transformation

A text document is represented by the occurrence of words it contains. The following approaches are used for document representation:

a. Bag of words

b. Vector spaces.

2.3 Feature Selection

It is also known as variable selection. Feature selection involves filtering important features for model creation. Feature selection is the subset of a more general field of feature extraction. Feature selection in text mining is mainly used in connection with applying known machine learning and statistical methods on text when addressing tasks such as Document Clustering or Document Classification. It can be categorized majorly into Filter, Wrapper, Embedded, and Hybrid methods.

The filter performs a statistical analysis over the feature space to select a discriminative subset of features. On the other hand Wrapper approach chooses a various subset of features that are first identified then evaluated using classifiers. While The embedded approach the feature selection process is embedded into the training phase of the classification. The hybrid approach takes advantage of both filter and wrapper approaches.

2.4 Text Mining Methods

At this point, Text mining becomes data mining including clustering, classification information retrieval, etc., can be used for text mining.

2.5 Interpretation/Evaluation

Analyzing the results and finding the new result by performing the above steps on several documents and texts.

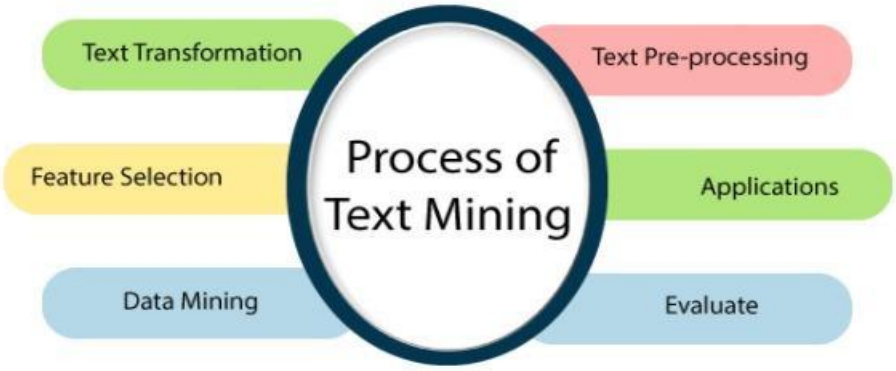

Figure 1: Process of Text Mining [16]

3. Areas of Text Mining

Text mining is an interdisciplinary field that incorporates areas such as information retrieval, information extraction, data mining, computational linguistics, and natural language processing. In today's increasing trend of publishing and storing scientific paper text mining is playing as one of the important and emerging technologies in evaluating many research papers and finding new information based on researches done by various individuals.

3.1 Information Extraction (IE)

The process of extracting structured information from unstructured and/or semi-structured text documents is known as Information Extraction. It does the pattern recognition in stored paper and maps the crucial and key information. IE systems can be used to extract abstract knowledge directly from a large volume of structured text sets or to extract concrete data from a set of documents which can then be further studied with classical data-mining techniques to discover more generalized patterns.

3.2 Information Retrieval (IR)

The algorithms used for representation, storage, and accessing of information items where the metadata handled is mostly in the form of textual documents, newspapers ad books, scientific papers which are retrieved from databases, and online platforms which contribute to research and journal. Information Retrieval is considered as an extension to document retrieval where the documents are processed to condense or extract the particular information needed to analyze research. An IR system allows us to narrow down the set of documents that are relevant to a particular research or researcher. One of the important platforms that use NLP and store and analyze the research paper is Google Scholarly.

3.3 Natural Language Processing (NLP)

Natural Language Processing is the most emerging concept in the field of artificial intelligence. It is defined as the study of human language so that computers can interpret natural languages similar to that of humans. Natural Language Processing is concerned with Natural Language Generation (NLG) and Natural Language Understanding (NLU). NLG ensures that generated text is grammatically and semantically correct. The NLG systems generally manage syntactic realizers to ensure that grammatical rules are followed and a text planner to decide how to manage sentences, paragraphs, and other parts clearly. The best Known NLG application is a machine translation.

3.4 Data Mining

Data Mining in scientific research papers refers to finding relevant information or discovering knowledge from large volumes of scientific papers. Data mining attempts to discover statistical rules and patterns automatically from the presented and published research scientific data. The motto of the data mining process is to extract information from a data set and transform it into an understandable structure for further analysis.

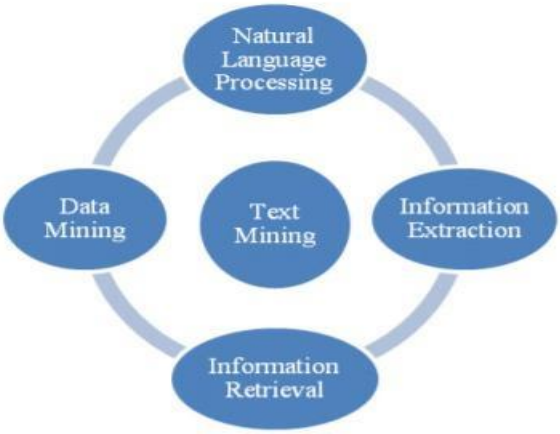

Figure 2: Application of text mining in the field of scientific research paper mining [16]

3.5 Sample Model for Illustrating Mining of Research Paper

An example of an Automatic Classifying and Ranking System for Scholastic Papers has been taken from the paper titled "Mining Scientific Articles powered by Machine

Learning Techniques" [1] for a broader understanding of the use of Text Mining and Machine Learning algorithms for mining research papers. The authors have classified the research papers using a Naive Bayes classification algorithm. Their model takes the word list as input from the trained model and the already classified model using a Naive Bayes Classifier. A ranking model is then generated, which can serve the purpose of a recommendation system for future searches.

Similarly, various other models can be created combining these powerful and insightful algorithms, and their efficiencies can be improved over time with the addition of the latest technologies and models. The main motivation should rely on the point of building a model which is accurate, less time-consuming, error-free and requires minimal or no human interventions, and still achieves the same results like that by a human expert.

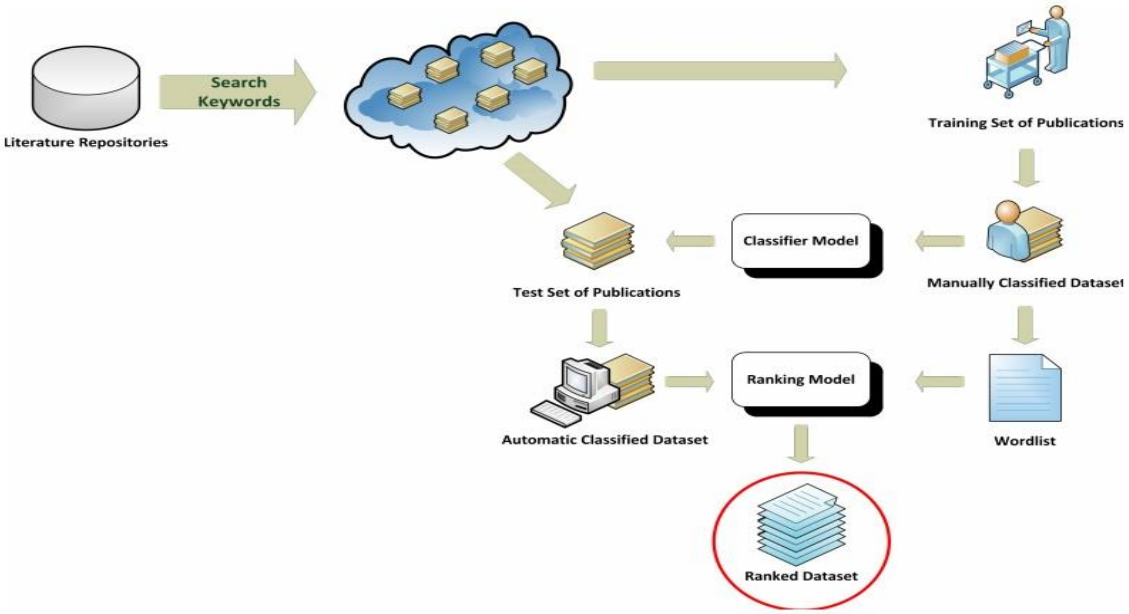

Figure 3: Architecture of a sample model process[1].

4. Tools and Platforms that can be used for Mining

4.1 NLTK

NLTK[6] is one of the top platforms for building programs in Python to work with human language data. It provides easy-to-use interfaces and it supports text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum. It is popular for teaching, learning, and working in computational linguistics with help of Python and is an incredible library to play with natural language.

Limitation

It's been mentioned that NLTK is "a complicated solution with an inflexible learning curve and a maze of internal limitations"[7]. For sentence tokenization, NLTK doesn't adhere to semantic analysis. Unlike Gensim, neural network models or word embeddings are absent in NLTK. It is slow, on the other hand, spaCy is said to be the fastest alternative. In fact, since NLTK was created for educational purposes, optimized runtime performance was never a goal. However, it's possible to accelerate execution using Python's multiprocessing module.

Matthew Honnibal, the creator of spaCy, noted that NLTK has numerous modules but very few (tokenization, stemming, visualization) are actually useful. Often NLTK leads to slow execution because of external libraries. The POS tagger was not good until Honnibal's averaged perceptron tagger was merged into NLTK in September 2015.

NLTK is evolving so fast that maintainers need to organize and arrange it often and throw away old things which are not needed.

4.2 spaCy

The motto of spaCy is to help you to work, build and gather real product and insight.[8]

The spaCy library has a time constraint and avoids wasting it. It is easy to use and install and its API is simple and productive. spaCy is good at extraction tasks for large-scale information. It is faster and efficient than NLTK.

Limitation

Spacy has a max_length limit of 1,000,000 characters. I was able to work on a document with 450,000 works flawlessly. However, the limit can be raised. It would split the text into chunks depending upon the size of the document.

4.3 ArXiv

It contains a large volume of scientific/ research papers and datasets of articles, making it possible to perform bibliometric studies not only on the metadata of papers but also their full-text content posted or published as e-prints after approval from moderators.[9] It is an open-source repository contributing to mining research papers, but it is not peer-reviewed. It consists of scientific papers in the fields of mathematics, physics, astronomy, electr ical engineering, computer science, quantitative biology, statistics, mathematical finance, and economics, which can be accessed online at its website[16].

Limitation

It does not peer-review papers and sometimes scientific papers are rejected for unknown reasons by moderators.

4.4 MALLET

MALLET is a Java-based software used for document classification, statistical natural language processing, information extraction, clustering, topic modeling, and other machine learning applications.[11]

MALLET includes advanced tools for the classification of documents. The software has efficient routines for constructing features from the text, it uses various machine learning algorithms (including Naïve Bayes, Maximum Entropy, and Decision Trees). Performance of various classifiers can be done using several commonly used metrics. [Quick Start] [Developer's Guide]

Apart from classification, MALLET includes facilities for sequence tagging for applications such as named-entity extraction from text. Algorithms include Maximum Entropy Markov Models, Hidden Markov Models, and Conditional Random Fields. These methods are implemented for finite-state transducers in an extensible system. [Quick Start] [Developer's Guide]

For analyzing large collections of unlabeled text, Topic models are used. The topic modeling tools contain efficient, sampling-based implementations of Hierarchical LDA and Latent Dirichlet Allocation. [Quick Start]

Numerical optimization is used in many of the algorithms in MALLET. Efficient implementation of Limited Memory BFGS can be done using MALLET. [Developer's Guide]

Apart from advanced Machine Learning applications, MALLET includes routines for making numerical representations for transforming text documents so that documents can be processed efficiently. The implementation uses the concept of "pipes" for handling distinct tasks such as removing stop words, tokenizing strings, and converting sequences into count vectors.

Limitation

The software cannot handle a very vast amount of data. MALLET is not intuitive. Their tools are really helpful, but one should go through their tutorials completely for efficient use.

4.5 AllenNLP

For careful and reproducible research, many use AllenNLP.[12] Instead of details, researchers can focus on the high-level summary of their models. It is widely adopted by Allen Institute for Artificial Intelligence. The library is widely used and has improved the quality of the research code. It gives knowledge about deep learning and makes it easier to share discoveries between teams. AllenNLP is very popular among the open-source community of contributors.

The AllenNLP team is committed to adding more features to this library in order to enable better research practices throughout the NLP community. It has a community of researchers who maintain a collection of the best models used in natural language processing.

Limitation:

It is not fully matured, not optimized for speed. **AllenNLP** takes too much time to process.

4.6 Genism

Genism is the fastest library for vector embedding and training.[14] The main algorithms in Gensim use highly optimized, battle-hardened, and parallel C routines. All Gensim source code is kept on Github and is under the GNU LGPL license. It is maintained by its open-source community. Business Support makes commercial arrangements. Using data-streamed algorithms, Gensim can process arbitrarily large corpora,. The software has no limitation of "dataset must fit in RAM ".

Gensim has an internal implementation limitation. it only shows 10,000 tokens per text item in their optimized paths. All additional tokens are ignored.

5. Feature Table for Various Tools

In this section, we have attempted to tabulate all the tools we have discussed above on some of the metrics like Features, Speed, Ease, etc. This will give a quick insight to researchers to compare different traits of all the tools available, as per their requirement.

Table 1. Comparison between different tools

S.No.	Tool	Features	Speed	Data Size	Ease of Use	Availability
1	NLTK	Tokenization, Stemming, tagging, parsing, and semantic reasoning	Slow	Supports big data	Complicated	Available for non commercial use, freely redistributed subject to license.
2	spaCy	Tokenization, part of speech tagging, named entity recognition. Supports 59+ languages	Fast as compared to NLTK	1,000,000 characters	Simple	Free and open-source library
3	ArXiv	e-prints	Fast	Limits to 10MB	Unknown	Open access repository
4	MALLET	Sequence tagging, name-entity extraction from text	Fast	Not good for big data	Easy to use but need training before diving in	Java package
5	AllenNLP	Framework for Deep Learning for NLP contains state-of-art reference models	Slow	Supports a wide variety of tasks and datasets	Simple	Open Source
6	Gensim	Fastest Library for training and vector embedding	Fast	Supports 10,000 tokens per item	Simple	Open Source

6. Conclusion

As from our findings, and the motivation behind our paper, we conclude that there are various tools and software present in the public domain which help in mining scholastic articles. Although, these tools/software cannot be simply compared apples-to-apples and declared as the sole winner, because of the different technologies and metrics they are based upon. But this can be seen as a feature, rather than a drawback.

From a normal researcher perspective, spaCy can be a very handy tool as it works upon tokenization, speech tagging and supports a variety of languages. Also, it is fast, super user-friendly, and is freely available to use. Similarly, if someone is working with huge datasets and machine learning, big data is involved, NLTK and AllenNLP can be of great assistance to them. Gensim is the fastest library for training and vector embedding. We have tried to list out various features of these tools who are the current market players, along with their limitations and drawbacks.

Thus, we see that depending upon the user requirements and technology stack they are working upon, the tools listed in our paper can be helpful and would ease the process of research. There is a lot of scope and voids currently in this area, and with the burgeoning research going on in this field, certainly, there would be more tools and libraries coming to our help in near future.

References

[1] Carlos A.S.J. Gulo, Thiago R.P.M. Rúbio, Shazia Tabassum, Simone G.D. Prado "Mining Scientific Articles powered by Machine Learning Techniques", Imperial College Computing Student Workshop (ICCSW 2015).

[2] Flynn, Louise Francis, and Matthew "Text mining handbook. In Casualty Actuarial Society", 2010.

[3] Rúbio, Carlos A.S.J. Gulo and Thiago R.P.M. "Text mining and scientific articles and using the R language. In Proceedings of the 10th Doctoral Symposium in Informatics Engineering", 2015.

[4] Iana Atanassova, Marc Bertin and Philipp Mayr "Mining Scientific Papers: NLP-enhanced Bibliometrics", 2019.

[5] EstelleChaix, LouiseDeléger, RobertBossy, ClaireNédellec "Text mining tools for extracting information about microbial biodiversity in food", 2019.

[6] "NLTK 3.6.2 documentation", https://www.nltk.org/

[7] "Natural Language Toolkit", https://devopedia.org/natural-language-toolkit

[8] "Industrial-Strength Natural Language Processing", https://spacy.io/

[9] https://stackoverflow.com/questions/48143769/spacy-nlp-library-what-is-maximum-reasonable-document-size#:~:text=Spacy%20has%20a%20max_length%20limit,chunks%20depending%20upon%20total%20size

[10] "arXiv", https://en.wikipedia.org/wiki/ArXiv

[11] "Machine Learning for Language Toolkit", http://mallet.cs.umass.edu/

[12] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi "AllenNLP: A Deep Semantic Natural Language Processing Platform"

[13] "Topic Modelling for Humans", https://radimrehurek.com/gensim/

[14] https://stackoverflow.com/questions/62543491/inconsistent-results-when-

training-gensim-model-with-gensim-downloader-vs-manual

[15] https://www.javatpoint.com/text-data-mining

[16] K.L.Sumathy, M.Chidambaram "Text Mining: Concepts, Applications, Tools and Issues – An Overview", International Journal of Computer Applications (0975 – 8887) Volume 80 – No.4, October 2013