Publisher homepage: www.universepg.com, ISSN: 2707-4625 (Online) & 2707-4617 (Print)
https://doi.org/10.34104/ijmms.023.034040
International Journal of Material and Mathematical Sciences
Journal homepage: www.universepg.com/journal/ijmms

Techniques, Methods, and Analysis of Text Mining: A Review

S. M. Abir Hasan[1]*, Mosharof Hossain[2], Md. Moniruzzaman[1], Md. Motasim Billah[1], Md. Sarowar Hossain Chowdhury[1], Ruhul Amin[3], Mst. Rakiba Parven[1], and Md. Ohedul Islam[1]

[1]Engineering Division, Bangladesh Atomic Energy Commission, Dhaka, Bangladesh; [2]Dept. of Statistics, Jahangirnagar University, Dhaka, Bangladesh; and [3]Institute of Nuclear Science and Technology, AERE, Bangladesh Atomic Energy Commission, Dhaka, Bangladesh.

*Correspondence: abir.sagar@yahoo.com (S. M. Abir Hasan, Engineering Division, Bangladesh Atomic Energy Commission (BAEC), Dhaka, Bangladesh).

ABSTRACT

Recent technological advances have led to the availability of new types of observations and measurements that were previously not available and that have fueled the 'Big Data' trend. Along with standard structured forms of data (containing mainly numbers), modern databases include new forms of unstructured data comprising words, images, sounds and videos which require new techniques to be exploited and interpreted. This study focuses on Text Mining, which is a set of statistical and computer science techniques specifically developed to analyze text data. New sources of text data are now available, such as text messaging, social media activity, blogs and web searches. The increasing availability of published text, sophisticated technologies and growing interest in organizations in extracting information from text have led to replacing (or at least supplementing) the human effort with automatic systems. Text mining can be used for a variety of scopes, ranging from basic descriptions of text content through word counts to more sophisticated uses such as finding links between authors and evaluating the content of scripts (e.g., automated marking of essays). Its basic purpose is to process the unstructured information contained in text data in order to make text accessible to various Data mining statistical algorithms. This could help make text data as informative as standard structured data and allow us to investigate relationships and patterns that would otherwise be extremely difficult, if not impossible, to discover. This study takes a quick look at how to organize and analyze unstructured text data using R programing language. And implementing various text mining operations to clean and structure the "eng_news_2020" dataset. This study also represents some association between the words using chi-square test and clustering procedure.

Keywords: Big data, Unstructured text data, R programming language, and Clustering procedure.

INTRODUCTION:

With the digital transformation of the entire world, there has been an explosion in textual information from a variety of sources. Text information refers to unstructured data such as HTML, XML, and document formats such as Microsoft Word, Adobe PDF, and email. Text mining is a type of data analysis that aims to retrieve valuable insights from textual information. It is part of the field of study referred to as the Natural Language Processing (NLP), which sits at the intersection of computational linguistics, computer science and artificial intelligence. NLP is a way for computers to analyze and understand human language. It is often used for machine interpretation, programmed address responding, and, of course, content mining. Text mining (Berry, 2004) is the computer- aided finding of previously undiscovered material through the automated extraction of information from a range of the

textual resources. An important factor is the association of the extracted information to form new facts or hypotheses that will be explored in more detail by more conventional experimental means. Text mining is different from what we know in web search. When searching, users often look for something that is already known and written by someone else. The point is to set aside all the material that doesn't currently meet your needs to find relevant information.

Text mining is a subset of data mining (Navathe *et al.,* 2000), which seeks for intriguing patterns in vast databases. Text mining, also known as intelligent text analysis, text data mining, or knowledge discovery in text (KDT), is a method that extracts information and knowledge from text. Interesting and non-trivial formulas from unstructured text Knowledge discovery from text (KDT) problems (Haralampos and Theodoulidis, 2001) use natural language processing (NLP) techniques to extract explicit and implicit concepts and semantic relationships between concepts. KDT is becoming increasingly relevant in new applications like as text comprehension.

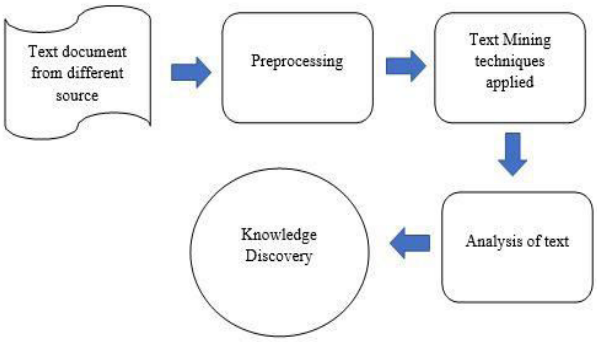


Fig. 1: Text Mining Process.

Fig. 1 represents the text mining process from 'Text document from the different source' to 'Knowledge Discovery'. The text mining method starts with gathering documents from various sources. A text mining tool retrieves a specific document and the preprocesses it by checking the format and character set. Next, the document will go through a text analysis phase. Text analysis is a semantic analysis to obtain quality information from text. There are many text analysis techniques; Depending on the organization's goals, a combination of techniques may be used. The information obtained can be fed into a management information system, creating a wealth of knowledge for the users of that system.

Methods and Models Used in Text Mining

Historically, a plethora of strategies have been created to address the challenge of text mining, which is basically collecting relevant information based on the user's needs. According on the information obtained, four major approaches appear to exist.

Term Based Method (TBM)

Term-based methods analyze documents based on terms and offer the benefits of efficient computational power and sophisticated term weighting theory. Polysemy and synonymy are issues that plague concept-based techniques (Salton and Buckley, 1988). Many of the found words' semantic meanings are unknown, making it difficult to satisfy user expectations. Many theoretically based solutions to this problem have been proposed via information retrieval.

Phrase Based Method (PBM)

Phrases contain more semantics, such as information, and are less ambiguous. Phrase-based methods analyze documents phrase by phrase because phrases are less ambiguous and discriminatory than individual terms (Ahonen *et al.,* 1998; Ali *et al*., 2021).

Concept Based Method (CBM)

At the sentence and document levels, terms are conceptually examined. Concept-based models are capable of distinguishing between irrelevant phrases and significant terms that explain the meaning of a sentence (Shehata *et al.,* 2007). Natural language processing methods are commonly used in concept-based models. Feature selection is applied to the query concepts to optimize the representation and eliminate noise and ambiguity.

Pattern Taxonomy Method (PTM)

Pattern taxonomy methods analyze documents based on patterns. Patterns can be structured into taxonomies using Is-A relationships. Pattern mining has been intensively researched in the data mining community for many years. Data mining approaches such as association rule mining, frequent item set mining, sequential pattern mining, and closed pattern mining can be used to uncover patterns (Wu *et al.,* 2004).

Techniques for Text Mining

There are the significant differences between human language and the computer language, but advances in technology are beginning to close this gap. Each of

these technologies plays an important role in text mining. The usefulness of each of these techniques varies depending on the situation.

Information Extraction (IE)

The starting point for computer analysis of unstructured text is information extraction. Information extraction software identifies key phrases and relationships in text. This is done by searching for predefined strings in the text. This is a process called pattern matching.

Topic Tracking

Topic tracking systems store user profiles and predict other documents that users may be interested in based on the documents they view. Yahoo offers a free topic tracking tool that allows users to select keywords and be notified when news about those topics becomes available.

Summarization

Text summarization can be very helpful in determining whether a large document meets your needs and is worth reading for more information. For large texts, text summarization software processes and summarizes the document in the time it takes a user to read the first paragraph. The key to summarizing is to shorten a document's length and content while retaining the important points and general meaning.

Categorization

Categorization involves identifying the main topics of a document by organizing it into a predefined set of topics. When classifying documents, computer programs often treat them as "groups of words". Unlike information extraction, it does not attempt to process actual information.

Clustering

Clustering could be a strategy for gathering comparable records together. However, the difference with classification is that documents are clustered on the fly and do not use predefined topics. Another benefit of clustering is that documents appear in multiple subtopics, which prevents useful documents from appearing in search results.

Textual Data Processing

Dataset Selection

For analysis purpose,

The “eng_news_2020” data set is selected from https://corpora.unileipzig.de/en?corpusId=eng_news_2020. The corpus “eng_news_2020” is an English news corpus based on material from 2020.

This includes 32,196,275 sets and 688,052,729 tokens. For not having enough computational power, smaller version of the dataset is chosen. The selected data set contains 10,000 sentences.

Tool Selection

There are many tools available for text data analysis. Most popular open-source tools are R and Python. R may be a dialect and environment for factual computing and design. It offers a variety of statistical and graphical techniques and is highly extensible. R is accessible as free program. It's easy to learn and use, and allows you to create well-designed, public- cation-quality plots. So, R console is used for this textual data analysis.

Pre-Processing

The raw text data collected is inherently unstructured for which it is necessary to clean it first. This preprocessing involves several steps described below.

Text Normalization

The data is converted into a standard format throughout this procedure. The "tm" library is used for all these tasks.

Tokenization

During this process, the complete content is partitioned into littler parts called tokens. The library "Tokenizer" can be used for tokenization.

Stemming

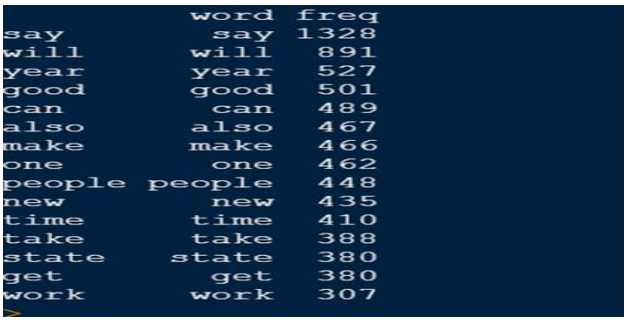
This is the process of separating words into stems or basic forms. Many text analysis packages (such as "quanteda" and "tm") in R employ "SnowballC" (Bouchet-Valat, 2014) to implement stemming, and it is now supported by 15 distinct languages.

Lemmatization

The reason for lemmatization compared to the stem method is to reduce inflectional forms to a common basic form. Lemmatization in R needs an extra software package, however stemming is generally adequate for languages with weak inflections, such as modern English.

Text Transformation

One of the most frequent forms for encoding a text corpus (i.e., a collection of texts) in a bag-of-word format is a document term matrix (DTM). A DTM is a matrix where the rows represent documents, the columns represent terms, and the cells represent the frequency of each term within each document. The advantage of this representation is that you can analyze the data using vector and matrix algebra, effectively moving from text to numbers. Furthermore, by using a special matrix format for sparse matrices, text data in DTM format is very memory efficient and can be analyzed with highly optimized operations (Kasper *et al.,* 2017).



	word	freq
say	say	1328
will	will	891
year	year	527
good	good	501
can	can	489
also	also	467
make	make	466
one	one	462
people	people	448
new	new	435
time	time	410
take	take	388
state	state	380
get	get	380
work	work	307

Fig. 2: Document Term Matrix (DTM).

Fig. 2 represents the Document Term Matrixx (DTM) which is actually processed text data, can be used for further investigation. How and what type of analysis should be done mainly depends on user & requirement.

Text Analysis

Text analysis involves extracting machine-readable facts from text. The goal of text analysis is to generate structured data from free text content. This process can be thought of as dividing large amounts of unstructured, disparate documents into data that is easier to manage and interpret. Text analysis is similar to other terms such as text mining, text analysis, and information extraction.

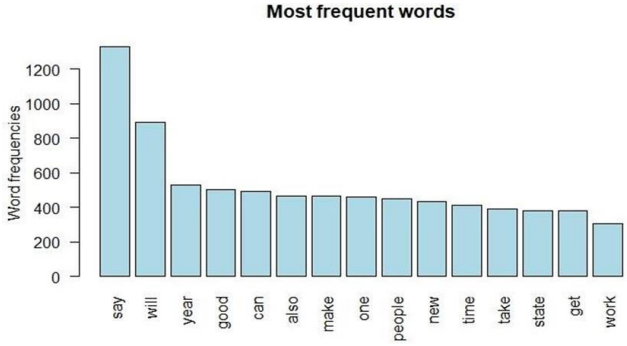


Fig. 3: Word Frequency Analysis (Bar Plot).

Fig. 3 represents the most frequent word in the text is "say" which occurred 1328 times. The next two frequent words are "will" and "year". This plot indicates about the most common used word in the news dataset.

Word Cloud

Another interesting way to observe word frequency is word cloud. A word cloud is a striking visual representation of "keywords" that commonly appear in text data. Rendering keywords creates a cloud-like color image, so you can quickly evaluate key text data.

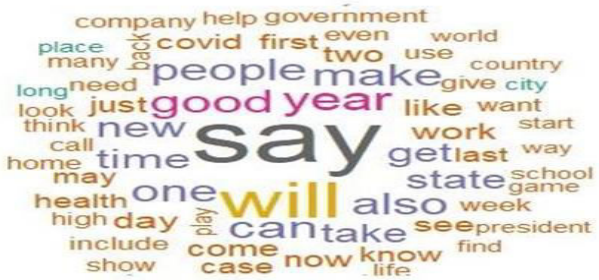


Fig. 4: Word Cloud.

Fig. 4 represents word clouds which are used to provide visual summaries of text; nevertheless, they have one important flaw: they lose track of the context. As a result, there is a danger of receiving erroneous and meaningless conclusions and making wrong assumptions about the underlying data.

Word Association

Word association is a form of content analysis of text data to find relationships between terms. In model relations, a special type of word association, the purpose is to calculate the similarity of candidate words' context texts, after collecting these contexts through a bag of words.

Pairs of very similar words can then be considered to have an exemplary relationship, meaning that these words share a common context. For phrase relationships, the fundamental concept is to count the amount of times two words appear together in a context. Correlation can be used effectively to analyze which words appear most frequently along with which words appear most frequently in survey responses, allowing one to see the context surrounding words.

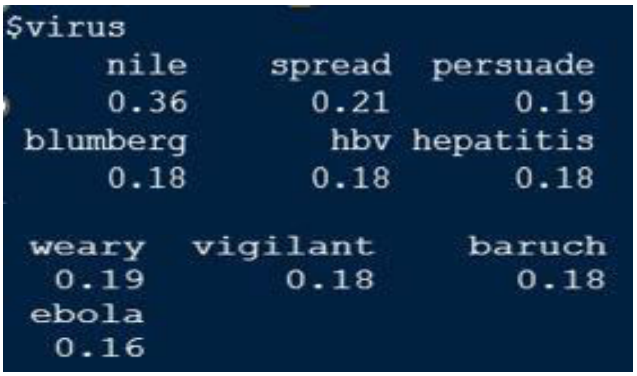


Fig. 5: Word Association.

Fig. 5 represents the word association where the output indicates that "nile" occurs 36% of the time with the word "virus" in the dataset. Again, "spread" occurs 21% of the time with the word "virus". So, it indicates that the words "virus" and "spread" are associated with each other but there is a very weak correlation between them.

Chi-Square Test

A chi-squared test (also chi-square or χ^2 test) is a statistical hypothesis test that is valid to perform when the test statistic is chi-squared distributed under the null hypothesis, specifically Pearson's chi-squared test and variants thereof. Pearson's chi-square test examines if there is a statistically significant discrepancy between the predicted and observed frequencies in a contingency table for one or more categories. This test is commonly used to classify observations into mutually exclusive groupsIf the null hypothesis is true, then there are no differences between classes in the population., then the test statistic derived from the observations follows a frequency distribution of χ^2.

The goal of the test is to assess how likely the observed frequency is, assuming the null hypothesis is true. A test statistic that follows a χ^2 distribution occurs when the observations are independent. There is also a χ^2 test to test the null hypothesis of pairwise independence based on observations of pairs of random variables. The chi-square test often refers to a test in which the distribution of the test statistic asymptotes to the χ^2 distribution. This means that the test statistic's sample distribution (if the null hypothesis is true) increasingly resembles a chi-squared distribution approximation as sample size increases.

A useful and simple procedure for detecting the statistical significance of the association of two qualitative variables in the 2 x 2 contingency table. A 2 x 2 contingency table is shown in **Table 1**.

Table 1: Contingency Table.

	Spread	No-Spread
Virus	60	225
No-Virus	26	9689

To test the hypotheses, the following steps are -

Step 1: Formulated of hypotheses:
The null hypothesis, H0: There are no relation between virus and spread.
The alternative hypothesis, H1: There are some relations between virus and spread.

Step 2: Level of significance:
Here significance level (assume), $\alpha = 0.05$.

Step 3: Selecting test statistic:
We need to use Chi-square statistic to test the null hypothesis. Under H0 the test statistic is:

$$\chi 2 = \sum_{k} \frac{(Observed - Expected)^2}{Expected}$$

Step 4: Finding the critical region:
The degrees of freedom: $v = (r-1)(c-1) = 1$
As, number of rows, $r = 2$
and number of columns, $c = 2$
The χ^2-value obtained from Statistical table for $\alpha = 0.05$ and $v = 1$ is 3.841. Hence, the critical region for right-tailed test is $\chi^2 > 3.841$.

Step 5: Computation of the statistic (from R):
X-squared=1378.7, df =1, p-value < 2.2e-16
From above it indicates that, χ^2-value calculated is higher than the tabulated value $\chi^2 = 1378.7 > 3.841$). And a p-value is near to zero ($2.2e-16$) and less than 0.05 (significant level). So, the null hypothesis can be rejected and concluded that virus and spread have some significant relations. This hypothesis test is based on the "eng_news_2020" corpus.

Word Network

Network text analysis involves the extraction and analysis of the networks from a text corpus. In these networks, nodes are concepts identified by words in the text, and edges between nodes represent relationships between concepts. Visualizing conceptual networks helps provide a compact representation of the general structure of the underlying text. Additionally, potential relationships between concepts that are not

explicit in the text are made visible. For example, approaches that visualize text as networks allow analysts to focus on important aspects without having to read a great deal of text (Paranyushkin, 2011).

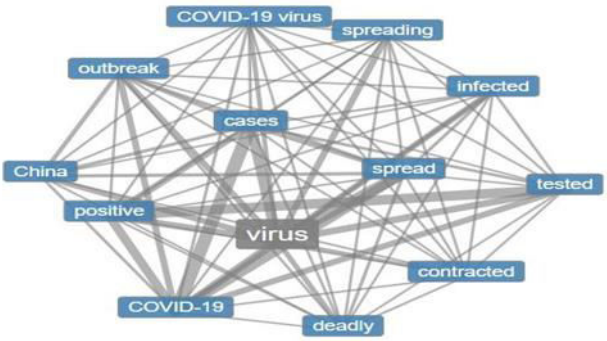


Fig. 6: Word Network.

From above word network, it is visualized that the word "virus" is related to "spread", "cases", "tested", "contracted", "deadly", "infected", "COVID-19", "positive", "China" and so on.

Simple Word Clusters

In hierarchical clustering, the clusters are repeatedly joined hierarchically, ending at a root. Hierarchical clustering works like a binary tree. Clustering methods that do not follow this principle are simply called flat clustering, but are also sometimes called nonhierarchical clustering or partitioned clustering. Hierarchical methods can be encourage partitioned into two subcategories. Agglomerative ("bottom-up") techniques start by organizing each object assigned to its own cluster and integrating them many times. A divisive approach ("top-down") has the opposite effect.

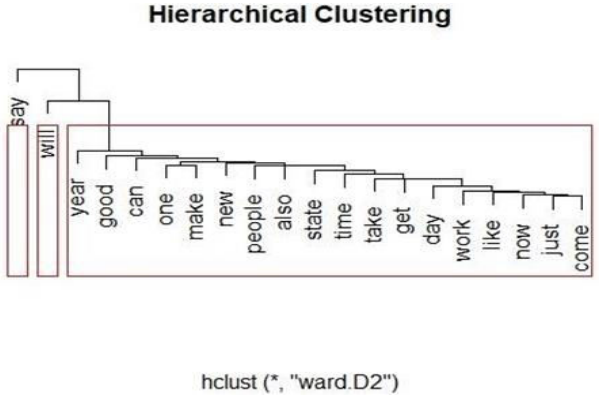


Fig. 7: Hierarchical Clustering (Dendrogram).

From above dendrogram, it is visualized that total number of words taken equal to 20 which are most frequent words. We used "ward method" to cluster these 20 words into 3 clusters. Cluster-1 only contains one word "say". Cluster-2 contains again one word which is "will". But the cluster-3 contains 18 words which are "year", "good", "can", "one", "make", "new", "people", "also", "state", "time", "take", "get", "day", "work", "like", "now", "just" and "come".

CONCLUSION:

The key advantage provided by text mining is the opportunity to exploit text records, on a very large scale. Here we have briefly described the techniques of text mining and procedure to implement text mining in a dataset. Text mining has a variety of potential applications in the field of education. In formative and summative assessment, for instance, it could be used to understand trends in vocabulary usage over time and the use of spelling and punctuation. To date, these applications have been completed by teachers and assessment experts without using advanced techniques such as text mining, but text mining allows the possibility of implementing these applications on a more comprehensive scale. The developments in NLP allow educational professionals to analyze the language structure of a vast amount of text documents in just a few minutes, plus the ongoing developments in this field could result in an increase in the accuracy of the findings.

The availability of novel data could lead, at least in principle, to novel measurement and research designs to address old and new research questions. However, working with very large, rich and new kind of datasets, it might not be straightforward to figure out what questions the data could answer accurately. Asking the right question might be more important now than ever (Einav and Levin, 2014).

Exploiting large text datasets without a proper research question might lead to a significant waste of resources. More heterogeneous and in-depth data could allow researchers to move from methods that allow the estimation of average relationships in the population towards differential effects for specific subpopulations of interest. This could mean looking at particular categories of students, defined by their specific background, level of achievement and other characteristics of interest. Text mining is an expanding field with the potential to support innovative areas of research. With careful research designs and proper

methods, text mining could make a salient contribution to educational research.

ACKNOWLEDGEMENT:

I would like to express our gratitude and thanks to our colleagues for their support and advice in completing this paper successfully.

CONFLICTS OF INTEREST:

The author (s) declares that there are no conflicts of interest to publish it.

REFERENCES:

1) Ahonen, H., Heinonen, O., and Verkamo, A. I. (1998). Applying data mining techniques for descriptive phrase extraction in digital document collections. In *Proceedings IEEE International Forum on Research and Technology Advances in Digital Libraries-ADL'98- IEEE*. (pp. 2-11). https://doi.org/10.1109/ADL.1998.670374
2) Ali MH, Hosain MS, and Hossain MA. (2021). Big data analysis using bigquery on cloud computing platform, *Aust. J. Eng. Innov. Technol*., **3**(1), 1-9. https://doi.org/10.34104/ajeit.021.0109
3) Berry Michael, W. (2004). Automatic discovery of similar words. Survey of Text Mining: Clustering, Classification and Retrieval". *Springer Verlag, New York, LLC*, 24-43. https://doi.org/10.1007/978-1-4757-4305-0_2
4) Bouchet-Valat, M. (2014). SnowballC: Snowball stemmers based on the C libstemmer UTF-8 library. *R package version,* **0.5**, 1. https://cran.r-project.org/web/packages/SnowballC/SnowballC.pdf
5) Einav, L., & Levin, J. (2014). The data revolution and economic analysis. *Innovation Policy and the Economy*, **14**(1), 1-24. https://doi.org/10.1086/674019
6) Karanikas, H., & Theodoulidis, B. (2002). Knowledge discovery in text and text mining software. *Centre for Research in Information Management, Department of Computation*.
7) Navathe, S. B., & Ramez, E. (2000). Data warehousing and data mining. *Fundamentals of Database Systems*, 841-872. https://books.google.com.bd/books/about/Fundamentals_of_Database_Systems.html?id=ZdhAQgAACAAJ&redir_esc=y
8) Paranyushkin, D. (2011). Identifying the pathways for meaning circulation using text network analysis. *Nodus Labs*, **26**, 1-26. https://noduslabs.com/wp-content/uploads/2012/04/Pathways-Meaning-Text-Network-Analysis.pdf
9) Salton, G., and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, **24**(5), 513-523. https://doi.org/10.1016/0306-4573(88)90021-0
10) Shehata, S., Karray, F., & Kamel, M. (2007). A concept-based model for enhancing text categorization. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 629-637. https://doi.org/10.1145/1281192.1281260
11) Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text analysis in R. *Communication methods and measures*, **11**(4), 245-265. https://doi.org/10.1080/19312458.2017.1387238
12) Wu, S. T., Xu, Y., & Chen, P. (2004, September). Automatic pattern-taxonomy extraction for web mining. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI'04), IEEE*, pp. 242- 248. https://doi.org/10.1109/WI.2004.10132

Citation: Hasan MAS, Hossain M, Moniruzzaman M, Billah MM, Chowdhury MSH, Amin R, Parven MR, and Islam MO. (2023). Techniques, methods, and analysis of text mining: a review, *Int. J. Mat. Math. Sci.*, **5**(5), 34-40. https://doi.org/10.34104/ijmms.023.034040