**Text Mining**

Text mining is one of the technologies of artificial intelligence that uses natural language processing to automatically extract data from different forms of text documents.

“The process of transforming unstructured textual data into structured data using natural language processing (NLP) which helps to get smart insights is known as **Text mining** ”.

Text mining is a field of data mining that analyzes a large volume of unstructured text data and extracts meaningful structured textual data. Text mining and Machine learning, combined concerns with automatic processing and analysis of unstructured textual data.

As the size of textual data is increasing exponentially, it is challenging to extract useful information from it. A huge collection of data is coming from social media, websites, institutes, databases, customer feedback, and e-mails which is unstructured text data. Organizing, categorizing, and capturing the information from raw data is a challenge to companies. Text mining allows one to gain meaningful insights from a wide range of data sources. It also helps to extract hidden patterns, rules, and relations from textual data.

The process of text mining involves a series of sub-tasks such as data gathering, pre-processing of data, data cleansing, and preparing data for NLP techniques to build input for machine learning models to perform text analysis.

**Text mining Techniques**

Text mining involves following techniques to extract the relevant information from raw data.

1. **Information Retrieval :-**

Information Retrieval mainly focuses on obtaining information from the source or database where information is particularly textual information. This system assists users to locate the required source of information and does not contribute to the overall solution of the problem stated. It will locate the location of storage which has required information stored in it. This system will give users a set of documents which are called “Relevant Documents”. Ideal Information Retrieval system will always return us “Relevant Documents”.

1. **NLP (Natural Language Processing) :-**

Natural language processing (NLP) performs the tasks with the help of machine learning and deep learning. It uses methods from various fields such as artificial intelligence, computer science, and data science to enable computers to understand the human language in written as well as verbal forms.

Computers cannot understand human language easily as there are many flaws and irregularities in patterns which humans speak. Hence there has to be set of techniques which must find patterns and then they can be used to understand patterns. This can be used to predict next to word to keep track of grammatical mistakes.

Amazon alexa, google’s assistant and apple’s siri these are some of the examples of day to day life entities which implement NLP in very efficient way.

1. **IE (Information Extraction) :-**

Whenever an organization is dealing with a huge amount of data, the task of collecting information becomes hectic and time consuming manually. Hence to automate the process and collect data automatically in lesser time information extraction techniques are implemented.

Formally information extraction is defined as “**Traversing through unstructured data and converting that data into more structured form is called information extraction.**”

Information extraction can be done by using deep learning and NLP algorithms like *‘named entity relation’.* These algorithms and techniques will make machines able to read and extract data from different sources like pdf, text files and other sources.

There has to be a different method for extracting data from different sources. For example if we are extracting data from the school result sheet then name, marks, percentage and other parameters should be extracted, and if there is extraction of data from the hotel bills then there will be different parameters for the extraction of data like order no. items, price per item and total amount.

1. **Data Mining :-**

Data mining is a concept which makes machines derive and extract useful information from the data given as input. In general cases data mining is considered to be related to numerical data. It is true in some sense but it does not mean that data mining cannot be applied to text or voice data. There is a huge amount of information stored in these formats on different platforms.

There are many techniques which are used to extract useful information from the text data, they are as follows:-

* 1. Sentiment analysis : -

Sentiment analysis means predicting and categorizing data into one the categories like positive, neutral or negative. This can be helpful for the large companies which are trying to divide the comments of customers on the product. This can also be used to categorize movie reviews given by viewers.

1. Text Summary :-

This involves breaking down complex and huge amounts of data into something which will be less in volume and more understandable. This can be breaking down complex words into more simpler ones or replacing large sentences with smaller phrases or words.

Along with above mentioned methods there are other application areas of data mining which can be explored as and when needed.

There are different methodologies which are used in the mentioned technique. They are as follows :

1. **Tokenization :-**

Tokenization as word suggests us is converting input text into smaller tokens or chunks. As it is hard for machines to have knowledge of Natural Language hence this tokenization makes the task of interpreting the textual data by machine using the mentioned method. In simple words it is used to convert paragraphs and text into smaller units called as tokens.

Consider an example: “I am learning natural language processing.” This single sentence has 6 words and without having a smaller format it will be difficult for machines to get the meaning of the sentence. Hence the above sentence can be divided into different words like ‘I’, ‘am’, ‘learning’, ‘natural’, ‘language’, ‘processing’.

Tokenization can be of different types. Depending on requirement and complexity of the problem, any of the below tokenization can be applied.

1. *Word Tokenization:-* In this type of tokenization sentence or given paragraph will be divided into different words. This tokenization can be done on the basis of simple space or based on any special symbol.

This tokenization technique can raise an issue when there is any out of vocabulary word in the given sentence. Hence tokenizer using which tokenization is to be done should be trained with a rich set of words. If the word is out of vocabulary then it will be replaced with unknown. But here arises a problem if there are n number of unknown then number of words can be same or there can be n different words which are unknown.

1. *Character tokenization :-* Character tokenization was introduced to overcome the issue raised by word tokenization. Here each sentence will be broken down into each of the characters of the word. Out of vocabulary issues in word tokenization will never be raised for each of the languages as the model only has to check entered characters with a standard set of characters in a given language. For instance, an English model will only investigate characters with 26 standard characters.

This type of tokenization comes with another challenge of keeping track of every character of the string. As for every word, the number of characters will increase in exponential order. There comes another issue of introducing  the meaning of words with characters.

1. *Sub - Word tokenization :-*

This tokenization method is similar to word tokenization. But the main idea here is tokenizing the word into suffixes and prefixes in the word. For example the sentence “I am one of the tallest people in the class”. This sentence can be tokenized into ‘I’, ‘am’, ‘one’, ‘of’, ‘the’, ‘tall’, ‘est’, ‘people’, ‘in’, ‘the’, ‘class’. Here we have made the words to be tokenized into words such that suffixes are also tokenized. This will be helpful when the model has out of vocabulary words. Just using suffixes can predict if a given word can be in the present tense or it can also be a noun.

1. **Stopwords :-**

Every language has a set of words which are commonly used in day to day use of language. Hence instead of focusing on these common words, focus can be taken to more important words in sentences.

For example, take the following sentence “Which active cricket player has the most number of centuries?” suppose we give this query to any search engine. So what should search engines do? Do they take each and every word and search them into the web? No, here search engines make use of the concept of stopwords and hence it will ignore the stopwords or words which are common in the given language. It will search for ‘player’, ‘most’, ‘centuries’ and it will have the required result as the name of the player.

Even though stop words are to be removed, there can be some cases where they don't need to be removed. For example sentences whose importance is to be determined by words like ‘not’, ‘their’ etc. Here we cannot afford to lose these stopwords.

1. **Stemming :-**

After tokenizing the given text document into tokens we have to reduce them to their root form. This is where stemming comes into picture. Let’s take an example of words like ‘runner’, ‘running’, ‘ran’ etc. to analyze all of these words will be a difficult task hence all of these words will be converted to their root form. This conversion to root form is called stemming. Most popular algorithm for stemming is PorterStemmer which is provided by the NLTK library of python.

One thing to notice here is words are not made to their root words based on their actual meaning when we make use of stemming technique. Sometimes they are just made to some root form just for the sake of making them to any other root form.

To overcome this drawback of stemming there is another method called Lematizing. Here words are made to their root form based on their true meaning in the English language.This lemmatizing is provided by WordNetLemmatizer in the NLTK library of python.

* 1. **Text Summarization**

Text summarization is a process of auto-generating a condensed version of useful text data that contains the main points that are useful to end users in the decision-making process. It allows generating a summary of text data from original unstructured data with conciseness and context.

Text summarization is used to summarize the text data, so it will be helpful to make appropriate decisions in less time. It is helpful in the sectors where large amounts of data should be analyzed to get a final decision.

It is used in newsletters, social media marketing, legal document summarization, financial research, etc

* 1. **Text Categorization:**

Text categorization is also known as Text classification. It is a form of supervised learning.

Text categorization is a text mining technique that uses natural language processing (NLP) to assign categories to unstructured data.

It is responsible for differentiating the text into predefined categories based on their content. It gathers all the text documents and analyzes them to classify the topics or categories.

Applications of text categorization are Language detection, Sentiment analysis, and Natural language inference.

Consider an example of language detection that helps to detect the language of text documents whether it is English, French, or Spanish and trains the model in that particular language. If the input is in another language then it will discard that message.

* 1. **Part-Of-Speech(POS) Tagging:**

Part of speech tagging is nothing but finding the part of speech of each word or token. It assigns a tag to a word denoting a part of speech- noun, pronoun, adjective, adverb, etc.

It takes a token as input and gives part of speech of that token as output. In POS Tagging converts the sentences into the list of words with their tag.

Consider an example, “I like to play football.” This sentence is converted into a list of words with their tag such as “I” as a preposition, “like” as a verb, “to” as to, “play” as a verb, and “football” as a noun.

[(“I”, “preposition”),(“like”, “verb”),(“to”, “to”),(“play”, “verb”),(“football”, “noun”)]

There are specified tags for each part of speech which are as follows :

|  |  |
| --- | --- |
| Noun (Singular) | NN |
| Noun (Plural) | NNS |
| Verb | VB |
| Adjective | JJ |
| Adverb | RB |
| Preposition | IN |

The list of words with tags is as follows

[(“I”, IN),(“like”, VB),(“to”, “to”),(“play”, VB),(“football”, NN)]

Note: Here “to” is a stopword that is removed during the process of Stopword removal.

* 1. **Sentiment Analysis :**

As the name suggests, sentiment means emotion or opinion, Sentiment analysis is also referred to as opinion analysis.

Sentiment analysis is the process of natural language processing(NLP) concerned with identifying and extracting the sentiments (i.e. positive, negative, neutral) of text data to track the customers' reviews over time.

It is primarily used to provide information about perceptions of brands, products, and services. It helps businesses to connect with customers and improve product quality and user experience.

Sentiment analysis can extract the polarity or the amount of positivity and negativity, subject and opinion holder within the text. This approach is used to analyze various parts of the text, such as a full document or a paragraph, sentence, or subsentence.

For example, consider an example of social media, There are many comments on social media posts to identify the positive and negative comments sentiment analysis can be used.

**Text Mining Tools and Technologies**

Text mining technologies encompass various techniques and tools used to extract insights and valuable information from textual data. These technologies include natural language processing (NLP) for language understanding, machine learning for training models on text, information retrieval for efficient searching, data visualization for visual representation of text, text summarization for condensing text, named entity recognition for identifying entities, topic modelling for discovering themes, sentiment analysis for determining sentiment, and text clustering for grouping similar texts. These technologies enable organizations to derive valuable insights, automate processes, and make data-driven decisions from textual data.

**Text Mining Tools**

Text mining tools refer to software applications or libraries that are designed to extract valuable information and insights from large volumes of text data. These tools utilize various techniques from natural language processing (NLP) and machine learning to analyse and process textual information.

**1.MonkeyLearn:**

A cloud-based text analysis software that uses machine learning and natural language processing (NLP) methods in order to extract important information from text data. For a variety of text mining tasks, it offers a user-friendly interface and pre-built models, helping users to perform complex text analysis task without any coding or machine learning experience.

By offering accessible and adaptable machine learning models, the platform aims to democratise text analysis by enabling users to extract useful information from their text data without the requirement for complex machine learning infrastructure or expertise.

**2.Google Cloud NLP:**

Natural language processing (NLP) capabilities are available through the Google Cloud Natural Language API, a cloud-based service offered by Google Cloud. By applying pre-trained models and powerful NLP algorithms, it enables developers to extract important information and insights from text.

Text is accepted as input by the Google Cloud Natural Language API, which then outputs several analyses. It is capable of carrying out operations including sentiment analysis, entity recognition, syntax analysis, content classification, and more.

The Google Cloud Natural Language API's underlying technology combines deep learning algorithms, trained models, and Google's vast data base. It keeps changing and getting better as

**3.IBM Watson:**

IBM Watson is a suite of AI-powered technologies and services developed by IBM. It leverages natural language processing (NLP), machine learning, and other advanced techniques to understand and analyse complex data, including text, images, and audio.

IBM Watson works by ingesting data and applying cognitive capabilities to extract insights and generate meaningful outcomes. It uses a combination of pre-trained models, custom models, and interactive interfaces to process and interpret data.

**4.Amazon Comprehend:**

Amazon Comprehend is a cloud-based natural language processing (NLP) service provided by Amazon Web Services (AWS). With the use of sophisticated NLP methods and pre-trained models, it enables developers to analyse text input and derive insightful information.

The service uses machine learning algorithms to understand the structure and meaning of text, extracting relevant information and generating structured output. It can handle text in multiple languages and provides high accuracy in analysing diverse types of textual content.

**5.Meaning Cloud:**

MeaningCloud is a cloud-based text analytics platform that offers a wide range of natural language processing (NLP) functionalities. It provides tools and services to analyse and extract meaningful insights from textual data. The platform uses a combination of rule-based approaches and statistical techniques to analyse and interpret text. It leverages machine learning algorithms to train models on large datasets, enabling accurate and efficient analysis of diverse text sources.

**6.Apache OpenNLP:**

Apache OpenNLP is an open-source library for natural language processing (NLP). It provides a set of tools and models to perform various NLP tasks, including tokenization, part-of-speech tagging, named entity recognition, parsing, and more. Apache OpenNLP works by employing statistical models and machine learning algorithms to analyze text data. It offers pre-trained models that have been trained on large annotated datasets to accurately perform linguistic analysis.

As an open-source project, Apache OpenNLP encourages community involvement and contributions, ensuring continuous improvement and the availability of new features.

**Applications of Text Mining**

Text mining also known as text analytics, has a wide range of applications across various industries. Right from academia and healthcare to businesses and social media platforms. This is giving rise to a number of text mining applications. Here are a few text mining applications used across the globe today:

1.**Sentiment analysis:**

Large amounts of text data can be mined for meaningful information using a technique called text mining. On the other side, sentiment analysis focuses on identifying the emotional tone or sentiment expressed in a text. Sentiment analysis may be carried out on a wide scale by utilising text mining techniques, allowing organisations and businesses to learn from client feedback, social media posts, reviews, and other sources of textual data. Then, using this data, decision-makers can better understand client sentiment, spot patterns, and develop new products and services.

**2.Content recommendation and personalization:**

Text mining is applicable in content recommendation by analysing and understanding the textual content of various sources such as articles, blogs, reviews, and social media posts. Through techniques like natural language processing and machine learning, text mining can extract relevant information and insights from the text. This information is then used to generate personalized content recommendations based on user preferences, interests, and previous interactions. By leveraging text mining, content recommendation systems can deliver more accurate and targeted suggestions to users, enhancing their browsing or shopping experience and increasing engagement.

**3.Fraud Detection:**

Text mining is applicable in fraud detection by analysing large volumes of text data, such as transaction records, customer reviews, emails, and online chats, to uncover patterns and anomalies that may indicate fraudulent activities. Text mining techniques, such as natural language processing and machine learning, can be used to extract relevant information, identify suspicious patterns, and detect fraudulent behaviour in textual data. By analysing text data, fraud detection systems can enhance their ability to identify potential fraudsters, detect fraudulent transactions or claims, and mitigate risks. Text mining enables organizations to efficiently process and analyse textual information, contributing to more accurate and effective fraud detection and prevention strategies.

**4.Document classification and categorization:**

Text mining is used in document classification and categorization to automatically analyse and classify large volumes of textual documents into predefined categories or classes. Through techniques like natural language processing and machine learning, text mining algorithms can extract features and patterns from the text, enabling accurate classification of documents based on their content, topics, or themes. This allows for efficient organization, retrieval, and management of documents, saving time

and improving information retrieval processes.

**5.Legal and e-discovery:**

Text mining is used in the legal industry and e-discovery processes to analyse and extract relevant information from large volumes of legal documents, contracts, court records, and electronic data. By employing natural language processing and machine learning techniques, text mining enables efficient search, retrieval, and analysis of legal text, aiding in case management, document review, and identifying patterns or insights relevant to legal investigations and proceedings. It helps legal professionals save time, improve accuracy, and gain valuable insights from complex and extensive legal document collections.

**6.Healthcare and biomedical research:**

Text mining is used in healthcare and biomedical research to analyse vast amounts of scientific literature, clinical records, and research papers. By employing natural language processing and machine learning techniques, text mining helps extract valuable information such as drug interactions, disease associations, treatment outcomes, and genetic patterns. It aids in knowledge discovery, literature review, and evidence-based decision-making, enabling researchers and healthcare professionals to gain insights, identify trends, and advance biomedical research and patient care.