Assignment No 9: Case Study on Text Mining

Aim:

case study on Text Mining

Theory:

Text mining, also known as text data mining, is the process of transforming unstructured text into a structured format to identify meaningful patterns and new insights. By applying advanced analytical techniques, such as Naïve Bayes, Support Vector Machines (SVM), and other deep learning algorithms, companies are able to explore and discover hidden relationships within their unstructured data.

Text is a one of the most common data types within databases. Depending on the database, this data can be organized as:

- **Structured data:** This data is standardized into a tabular format with numerous rows and columns, making it easier to store and process for analysis and machine learning algorithms. Structured data can include inputs such as names, addresses, and phone numbers.
- **Unstructured data:** This data does not have a predefined data format. It can include text from sources, like social media or product reviews, or rich media formats like, video and audio files.
- **Semi-structured data:** As the name suggests, this data is a blend between structured and unstructured data formats. While it has some organization, it doesn’t have enough structure to meet the requirements of a relational database. Examples of semi-structured data include XML, JSON and HTML files.

Text mining techniques

The process of text mining comprises several activities that enable you to deduce information from unstructured text data. Before you can apply different text mining techniques, you must start with text preprocessing, which is the practice of cleaning and transforming text data into a usable format. This practice is a core aspect of natural language processing (NLP) and it usually involves the use of techniques such as language identification, tokenization, part-of-speech tagging, chunking, and syntax parsing to format data appropriately for analysis. When text preprocessing is complete, you can apply text mining algorithms to derive insights from the data. Some of these common text mining techniques include:

1. **Information retrieval**

Information retrieval (IR) returns relevant information or documents based on a pre-defined set of queries or phrases. IR systems utilize algorithms to track user behaviors and identify relevant data. Information retrieval is commonly used in library catalogue systems and popular search engines, like Google. Some common IR sub-tasks include:

- Tokenization: This is the process of breaking out long-form text into sentences and words called “tokens”. These are, then, used in the models, like bag-of-words, for text clustering and document matching tasks.

- Stemming: This refers to the process of separating the prefixes and suffixes from words to derive the root word form and meaning. This technique improves information retrieval by reducing the size of indexing files.

2. Natural language processing (NLP)

Natural language processing, which evolved from computational linguistics, uses methods from various disciplines, such as computer science, artificial intelligence, linguistics, and data science, to enable computers to understand human language in both written and verbal forms. By analyzing sentence structure and grammar, NLP sub-tasks allow computers to “read”. Common sub-tasks include:

- Summarization: This technique provides a synopsis of long pieces of text to create a concise, coherent summary of a document’s main points.

- Part-of-Speech (PoS) tagging: This technique assigns a tag to every token in a document based on its part of speech—i.e. denoting nouns, verbs, adjectives, etc. This step enables semantic analysis on unstructured text.

- Text categorization: This task, which is also known as text classification, is responsible for analyzing text documents and classifying them based on predefined topics or categories. This sub-task is particularly helpful when categorizing synonyms and abbreviations.

- Sentiment analysis: This task detects positive or negative sentiment from internal or external data sources, allowing you to track changes in customer attitudes over time. It is commonly used to provide information about perceptions of brands, products, and services. These insights can propel businesses to connect with customers and improve processes and user experiences.

3. Information extraction

Information extraction (IE) surfaces the relevant pieces of data when searching various documents. It also focuses on extracting structured information from free text and storing these entities, attributes, and relationship information in a database. Common information extraction sub-tasks include:

- Feature selection, or attribute selection, is the process of selecting the important features (dimensions) to contribute the most to output of a predictive analytics model.

- Feature extraction is the process of selecting a subset of features to improve the accuracy of a classification task. This is particularly important for dimensionality reduction.

- Named-entity recognition (NER) also known as entity identification or entity extraction, aims to find and categorize specific entities in text, such as names or locations. For example, NER identifies “California” as a location and “Mary” as a woman’s name.

Text mining applications:

1. Customer services
2. Risk Management
3. Maintenance
4. Health Care
5. Spam Filtering

Conclusion:

Text mining is a powerful tool for extracting insights from large datasets of text data such as customer feedback and reviews. By leveraging natural language processing and machine learning algorithms, companies can gain a deeper understanding of their customers' needs and preferences, and make data-driven decisions to improve their products and services.