MET415 Introduction to Business Analytics
Module 3

Database Management System (DBMS)

DBMS is a traditional old system. A Database Management System (DBMS) is a software system that is designed to manage and organize data in a structured manner. It allows users to create, modify, and query a database, as well as manage the security and access controls for that database.

- For example, DBMS of College having tables of data for students, faculty, courses, grades etc.
- DBMS can store data in GBs (small storage)
- DBMS is meant for specific purpose
- Scope of performing Analytics in DBMS is very limited

To store large amount of data and to perform Analytics effectively, Data warehouses are required.

Data Warehouse

A Data Warehouse is separate from DBMS, it stores a huge amount of data, which is typically collected from multiple heterogeneous sources like files, DBMS, etc. The goal is to produce statistical results that may help in decision makings.

- For example, a college might want to see quick different results, like how the placement of CS students has improved over the last 10 years, in terms of salaries, counts, etc.

To effectively perform analytics, an organization keeps a central Data Warehouse to closely study its business by organizing, understanding, and using its historic data for taking strategic decisions and analyzing trends.

Benefits of Datawarehouse

1. **Better business analytics:** Data warehouse plays an important role in every business to store and analysis of all the past data and records of the company. which can further increase the understanding or analysis of data to the company.
2. **Faster Queries:** Data warehouse is designed to handle large queries that's why it runs queries faster than the database.
3. **Improved data Quality:** In the data warehouse the data you gathered from different sources is being stored and analyzed it does not interfere with or add data by itself so your quality of data is maintained and if you get any issue regarding data quality then the data warehouse team will solve this.
4. **Historical Insight:** The warehouse stores all your historical data which contains details about the business so that one can analyze it at any time and extract insights from it

Different between DBMS and Data warehouse

S.No.	Database	Data Warehouse
1.	A common Database is based on operational or transactional processing. Each operation is an indivisible transaction.	A Data Warehouse is based on analytical processing.
2.	Generally, a Database stores current and up-to-date data which is used for daily operations.	A Data Warehouse maintains historical data over time. Historical data is the data kept over years and can be used for trend analysis, make future predictions and decision support.
3.	A database is generally application specific. Example - A database stores related data, such as the student details in a school.	A Data Warehouse is integrated generally at the organization level, by combining data from different databases. Example - A data warehouse integrates the data from one or more databases, so that analysis can be done to get results, such as the best performing school in a city.
4.	Constructing a Database is not so expensive.	Constructing a Data Warehouse can be expensive.

DATABASE	DATA WAREHOUSE
Normalized data structure is there in a database in separate tables	Denormalized data structure is used for enhanced analytical response time
Transactional function is carried out, though analytic is possible but are difficult to perform due to complexity of normalized data	Dynamic and quick analysis of data is done

Example Applications of Data Warehousing
Data Warehousing can be applied anywhere where we have a huge amount of data and we want to see statistical results that help in decision making.

Social Media Websites: The social networking websites like Facebook, Twitter, Linkedin, etc. are based on analyzing large data sets. These sites gather data related to members, groups, locations, etc., and store it in a single central repository. Being a large amount of data, Data Warehouse is needed for implementing the same.

Banking: Most of the banks these days use warehouses to see the spending patterns of account/cardholders. They use this to provide them with special offers, deals, etc.

Government: Government uses a data warehouse to store and analyze tax payments which are used to detect tax thefts.

Characteristics of Data Warehouse

The characteristics of Data Warehouse are

Centralized Data Repository: Data warehousing provides a centralized repository for all enterprise data from various sources, such as transactional databases, operational systems, and external sources. This enables organizations to have a comprehensive view of their data, which can help in making informed business decisions.

Data Integration: Data warehousing integrates data from different sources into a single, unified view, which can help in eliminating data silos and reducing data inconsistencies.

Historical Data Storage: Data warehousing stores historical data, which enables organizations to analyze data trends over time. This can help in identifying patterns and anomalies in the data, which can be used to improve business performance.

Query and Analysis: Data warehousing provides powerful query and analysis capabilities that enable users to explore and analyze data in different ways. This can help in identifying patterns and trends, and can also help in making informed business decisions.

Data Transformation: Data warehousing includes a process of data transformation, which involves cleaning, filtering, and formatting data from various sources to make it consistent and usable. This can help in improving data quality and reducing data inconsistencies.

Data Mining: Data warehousing provides data mining capabilities, which enable organizations to discover hidden patterns and relationships in their data. This can help in identifying new opportunities, predicting future trends, and mitigating risks.

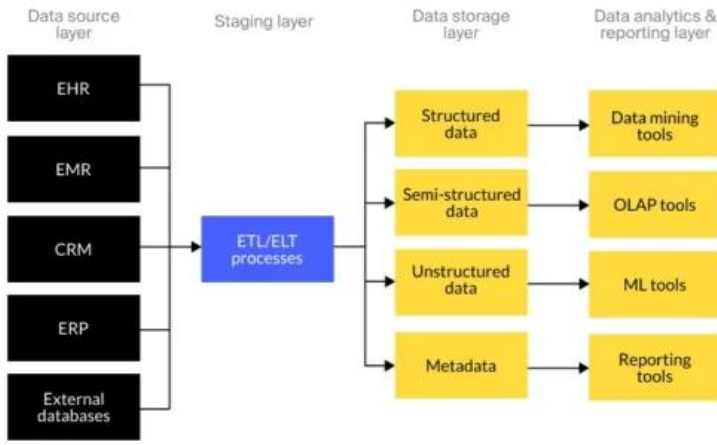
Data Security: Data warehousing provides robust data security features, such as access controls, data encryption, and data backups, which ensure that the data is secure and protected from unauthorized access.

Design Considerations of Data Warehouse

Design considerations of data warehouse involves designing data warehouse architecture.

Data Warehouse Architecture typically consists of four core components

1. data source layer
2. data staging
3. data storage and
4. data analytics and Reporting layer



Data source layer aggregates healthcare data from internal and external sources like electronic health records (EHR), electronic medical records (EMR), claims management systems, public health registries, pharmacy management systems, etc.)

Data staging layer serves as a temporary storage where data undergoes the extract, transform and load (ETL) or extract, load and transform (ELT) processes so that only relevant and useful data is transferred further into the warehouse

Data storage layer is a central database that can also encompass data marts to serve different lines of business (HR, accounting, etc.) or departments (radiology, pediatrics, etc.)

Data analytics layer comprises business intelligence, online analytical processing (OLAP) and data analytics systems as well as data reporting and visualization tools.

when designing a data warehouse, the primary factor to be taken into consideration is the needs of end users

Design Consideration of a Data Warehouse for heathcare

Streamlining data integration
Healthcare data sources provide petabytes-worth of information — patient, financial, pharmaceutical, clinical research, as well as data from wearables and IoT devices. This data comes in a variety of forms and formats, from structured, numeric data to unstructured written texts, radiology images, and more. One approach to power efficient data extraction and consolidation is to use semantic technologies that enrich data by adding context and facilitate more meaningful integration.

Designing data models
The data model is the backbone of a data warehouse and it has a significant effect on time-to-value and system adaptability. Basically, there are two approaches — top-down and bottom-up.

In a top-down model, “atomic” data is stored at the lowest level of granularity. Data elements are assigned specific categories and dimensions that will be represented in a schema, hence this approach requires a very detailed specification of how the data will be used. Although this model is comprehensive, it is very inflexible if new data that needs to enter the system does not fit into the predefined categories and dimensions.

A bottom-up model suggests creating smaller data marts to address specific business or clinical needs. These data marts pull information from a data superset into individual applications tailored to the needs of specific departments or areas of research. This is a cost-effective method to start a data warehousing project but the drawback is that large-scale analysis or reporting across these data marts is extremely difficult.

Ensuring security and privacy

Of all personal data, health data is considered most sensitive. In addition to routine clinical information like medical history and lab results, health information systems contain names, addresses, health insurance card numbers, and other confidential information. Any security flaw in a healthcare data warehouse could lead to unauthorized access and threaten patient privacy.

To keep protected health information (PHI) secure and private, data warehouse development requires a security-first approach based on security best practices and policies like data encryption and data pseudonymization for patient de-identification. Other security measures include multi-factor authorization and granular access control with custom roles and permissions.

In addition, any data storage that handles health records must comply with the Health Insurance Portability and Accountability Act (HIPAA). Hence, when building a cloud-based healthcare data warehouse, it's important to choose a HIPAA-compliant cloud storage service provider.

Optimizing performance

Since a healthcare data warehouse underpin clinical decision-making, glitch-free performance is imperative. This is however no easy task to achieve since a data warehouse processes

diverse sets of data and answers ad hoc queries from multiple users.

To ensure fast retrieval of data and reduce response time, best practices include leveraging materialized view support, bitmap indexing, result caching, and optimized queries, among other things. Equally important for reliable performance is to ensure automated and cost-effective scaling of compute resources to meet the changing analytics needs.

Data Lakes

A data lake is a central location that holds a large amount of data in its native, raw format. Compared to a hierarchical data warehouse, which stores data in files or folders, a data lake uses a flat architecture and object storage to store the data.

First and foremost, data lakes are open format, so users avoid lock-in to a proprietary system like a data warehouse, which has become increasingly important in modern data architectures.

Data lakes are also highly durable and low cost, because of their ability to scale and leverage object storage.

Basic Characteristics of Data Lake

1. **Data Fidelity:** A data lake stores data as it is in a business system. Different from a data warehouse, a data lake stores raw data, whose format, schema, and content cannot be modified. A data lake stores your business data as it is. The stored data can include data of any format and of any type.

2. **Data Flexibility:** Schema-on-write or schema-on-read are two types of Schema which indicates the phase in which the data schema is designed. A schema is essential for any data

application. Schema-on-write means a schema for data importing is determined based on a specific business access mode before data is written. This enables effective adaptation between data and your businesses but increases the cost of data warehouse maintenance at the early stage.

A data lake adopts schema-on-read, meaning it sees business uncertainty as the norm and can adapt to unpredictable business changes. You can design a data schema in any phase as needed, so the entire infrastructure generates data that meets your business needs. Fidelity and flexibility are closely related to each other. Since business changes are unpredictable, you can always keep data as-is and process data as needed. Therefore, a data lake is more suitable for innovative enterprises and enterprises with rapid business changes and growth. A data lake is intended for data scientists and business analysts that usually need highly efficient data processing and analytics and prefer to use visual tools.

3. **Data Manageability:** A data lake provides comprehensive data management capabilities. Due to its fidelity and flexibility, a data lake stores at least two types of data: raw data and processed data. The stored data constantly accumulates and evolves. This requires robust data management capabilities, which cover data sources, data connections, data formats, and data schemas. A data schema includes a database and related tables, columns, and rows. A data lake provides centralized storage for the data of an enterprise or organization. This requires permission management capabilities.

4. **Data Traceability:** A data lake stores the full data of an organization or enterprise and manages the stored data throughout its lifecycle, from data definition, access, and storage to processing, analytics, and application. A robust data lake fully reproduces the data production process and

data flow, ensuring that each data record is traceable through the processes of access, storage, processing, and consumption.

A data lake requires a wide range of computing capabilities to meet your business needs.

5. Data Rich Computing Engines: A data lake supports a diversity of computing engines, including batch processing, stream computing, interactive analytics, and machine learning engines. Batch processing engines are used for data loading, conversion, and processing. Stream computing engines are uses for real-time computing. Interactive analytics engines are used for exploratory analytics. The combination of big data and artificial intelligence (AI) gave birth to a variety of machine learning and deep learning algorithms. For example, TensorFlow and PyTorch can be trained on sample data from the Hadoop Distributed File System (HDFS), Amazon S3, or Alibaba Cloud Object Storage Service (OSS). Therefore, a qualified data lake project should provide support for scalable and pluggable computing engines.

6. Multi-Modal Storage Engine: In theory, a data lake should provide a built-in multi-modal storage engine to enable data access by different applications, while considering a series of factors, such as the response time (RT), concurrency, access frequency, and costs. However, in reality, the data stored in a data lake is not frequently accessed, and data lake-related applications are still in the exploration stage. To strike a balance between cost and performance, a data lake is typically built by using relatively inexpensive storage engines, such as Amazon S3, Alibaba Cloud OSS, HDFS, or Object-Based Storage (OBS).

Data Mining

Data mining is the process of detecting anomalies, patterns, and correlations within massive databases to forecast future results. This is accomplished by combining three intertwined fields: statistics, artificial intelligence, and machine learning. Data mining is simply sorting through data to find something valuable.

Example: Mining, on a smaller scale, is an activity that involves gathering data in one location in some structure. For example, creating an Excel spreadsheet or summarizing the main points of a text.

Data mining is all about:

- processing data;
- extracting relevant and valuable insights out of it

Data Mining Process

The data mining process is divided into five process steps.

Learning more about each process step helps explain how data mining works.

Collection. Data is stored, organized, and loaded into a data warehouse. The data is stored and handled on in-house servers or in the cloud.

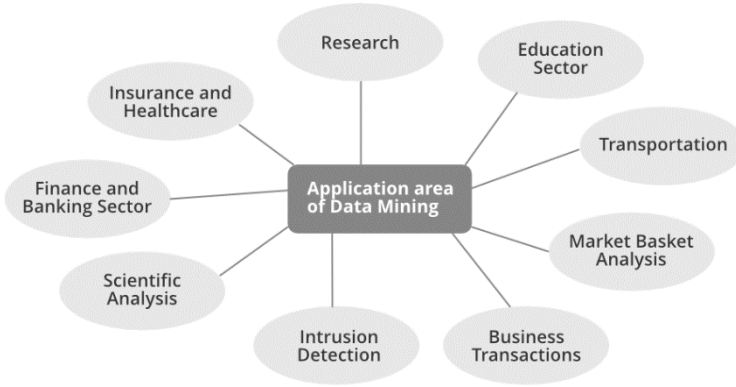
Understanding. Data Scientists and Business Analysts will inspect the “gross” or “surface” properties of the data. After that, they conduct a more in-depth analysis from the perspective of a problem statement defined by the business. This can be covered using querying, reporting, and visualization.

Preparation. After confirming the availability of data sources, they must be cleaned, constructed, and formatted into the desired form. This stage may include more in-depth data exploration based on the insights uncovered in the former stage.

Modeling. At this stage, modeling techniques for the prepared dataset are chosen. A data model is a visual representation of the relationships between several types of information stored in a database. A sales transaction, for example, is divided into related data points that describe the customer, the seller, the object sold, and the payment methods. Each object must be described in detail to be accurately stored and retrieved from a database.

Evaluation. Finally, the model results are analyzed concerning business objectives. New business requirements may be raised throughout this phase due to new patterns discovered in the model results or other factors.

Data Mining Applications



Different Applications of Data Mining are

Scientific Analysis: Scientific simulations are generating bulks of data every day. This includes data collected from nuclear laboratories, data about human psychology, etc. Data mining techniques are capable of the analysis of these data. Now we can capture and store more new data faster than we can analyze the old data already accumulated. Example of scientific analysis:

- Sequence analysis in bioinformatics
- Classification of astronomical objects
- Medical decision support.

Intrusion Detection: A network intrusion refers to any unauthorized activity on a digital network. Network intrusions often involve stealing valuable network resources. Data mining technique plays a vital role in searching intrusion detection, network attacks, and anomalies. These techniques help in selecting and refining useful and relevant information from large data sets. Data mining technique helps in classify relevant data for Intrusion Detection System. Intrusion Detection system generates alarms for the network traffic about the foreign invasions in the system. For example:

- Detect security violations
- Misuse Detection
- Anomaly Detection

Business Transactions: Every business industry has data stored in memory for ever. Such transactions are usually time-related and can be inter-business deals or intra-business operations. The effective and in-time use of the data in a reasonable time frame for competitive decision-making is definitely the most important problem to solve for businesses. Data mining helps to analyze these business transactions and identify marketing approaches and decision-making.
For example

- Direct mail targeting

- Stock trading
- Customer segmentation
- Churn prediction (Churn prediction is one of the most popular Big Data use cases in business)

Market Basket Analysis: Market Basket Analysis is a technique that gives the careful study of purchases done by a customer in a supermarket. This concept identifies the pattern of frequent purchase items by customers. This analysis can help to promote deals, offers, sale by the companies and data mining techniques helps to achieve this analysis task. Example:

- Data mining concepts are in use for Sales and marketing to provide better customer service, to improve cross-selling opportunities, to increase direct mail response rates.
- Customer Retention in the form of pattern identification and prediction of likely defections is possible by Data mining.
- Risk Assessment and Fraud area also use the data-mining concept for identifying inappropriate or unusual behavior etc.

Education: For analyzing the education sector, data mining uses Educational Data Mining (EDM) method. This method generates patterns that can be used both by learners and educators. By using data mining EDM we can perform some educational task:

- Predicting students admission in higher education
- Predicting students profiling
- Predicting student performance
- Teachers teaching performance
- Curriculum development
- Predicting student placement opportunities

Research: A data mining technique can perform predictions, classification, clustering, associations, and grouping of data with perfection in the research area. In most of the technical research in data mining, we create a training model and testing model. The training/testing model is a strategy to measure the precision of the proposed model. It is called Train/Test because we split the data set into two sets: a training data set and a testing data set. A training data set used to design the training model whereas testing data set is used in the testing model.

For Example:

- Classification of uncertain data.
- Information-based clustering.
- Decision support system
- Web Mining
- Domain-driven data mining
- IoT (Internet of Things) and Cybersecurity
- Smart farming IoT (Internet of Things)

Healthcare and Insurance: A Pharmaceutical sector can examine its new deals force activity and their outcomes to improve the focusing of high-value physicians and figure out which promoting activities will have the best effect in the following upcoming months, Whereas the Insurance sector, data mining can help to predict which customers will buy new policies, identify behavior patterns of risky customers and identify fraudulent behavior of customers.

- Claims analysis i.e which medical procedures are claimed together.
- Identify successful medical therapies for different illnesses.
- Characterizes patient behavior to predict office visits.

Transportation: A diversified transportation company with a large direct sales force can apply data mining to identify the best prospects for its services. A large consumer merchandise organization can apply information mining to improve its business cycle to retailers.

- Determine the distribution schedules among outlets.
- Analyze loading patterns.

Financial/Banking Sector: A credit card company can leverage its vast warehouse of customer transaction data to identify customers most likely to be interested in a new credit product.

- Credit card fraud detection.
- Identify 'Loyal' customers.
- Extraction of information related to customers.
- Determine credit card spending by customer groups.

Software Tools used in Data Mining

Most popular software tools used for Data Mining is listed here.

- **Monkey Learn** | No-code text mining tools
- **RapidMiner** | Drag and drop workflows or data mining in Python
- **Oracle Data Mining** | Predictive data mining models
- **IBM SPSS Modeler** | A predictive analytics platform for data scientists
- **Weka** | Open-source software for data mining
- **Knime** | pre-built components for data mining projects
- **H2O** | Open-source library offering data mining in Python
- **Orange** | Open-source data mining toolbox

- **Apache Mahout** | Ideal for complex and large-scale data mining
- **SAS Enterprise Miner** | Solve business problems with data mining

Text Analytics (Text Mining)

Text analytics is the automated process of translating large volumes of unstructured text into quantitative data to uncover insights, trends, and patterns. Combined with data visualization tools, this technique enables companies to understand the story behind the numbers and make better decisions.

Text mining, text analysis, and text analytics are often used interchangeably, with the end goal of analyzing unstructured text to obtain insights. However, while text mining (or text analysis) provides insights of a qualitative nature, text analytics aggregates these results and turns them into something that can be quantified and visualized through charts and reports.

Real-world applications of Text mining:

1. Fraud Detection: Text analytics, (text mining techniques), presents a huge potential for domains that collect the entirety of their data in text format. This is an opportunity that insurance and finance businesses are seizing. These businesses are now capable of processing claims quickly as well as preventing and detecting fraud by integrating the results of text analysis with pertinent structured data.

2. Social Media Analysis: Many text mining methods have been developed specifically for assessing the performance of social media sites. These aid with the tracking and

interpretation of online content created by the news, blogs, emails, and other sources. Text mining technologies can also quickly evaluate the number of posts, likes, and connections your brand has on social media, helping you to better understand how people are reacting to your company and online content. The research will help you figure out “what’s popular and what isn’t” for your target market.

3. Customer Care Service: In the realm of customer service, text mining applications, notably natural language processing (NLP), are becoming increasingly important. Companies are adopting text analytics tools to improve their entire customer experience by gaining access to textual data from a variety of sources, including surveys, user feedback, and user conversations, among others. Text analysis seeks to minimize the company’s reaction time and assist in quickly and efficiently resolving client complaints.

4. Knowledge Management: Managing a large volume of text data has become a difficulty in several areas, such as healthcare. It would certainly fly to the moon if you began designing systems and kept all of the papers relevant to healthcare on a single upwardly scalable rack. The amount of data collected per hour is enormous. All of this information must be kept in such a way that it may be accessed whenever it is needed. It’s possible that an epidemic could break out, and hospitals will need to work together to analyze all of their data in order to locate the source or the first affected individual.

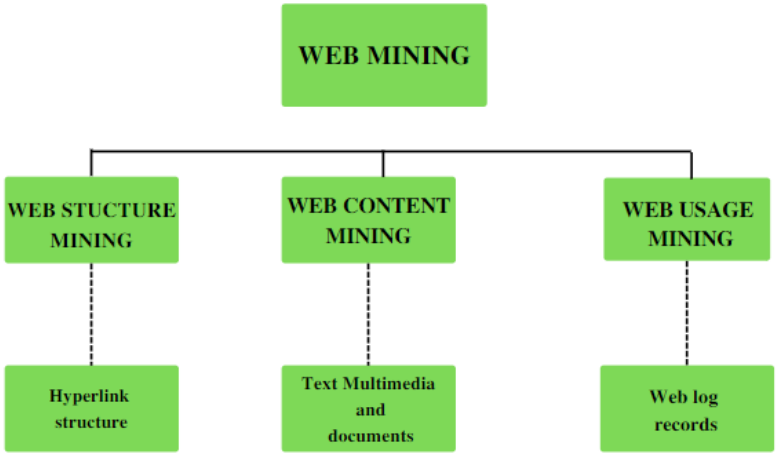
5. Risk Management: The absence of appropriate or inadequate risk analysis is one of the leading reasons for company failure. Implementing and adopting risk management tools based on text mining technologies, like SAS Text Miner, may assist organizations in staying current with industry trends and enhancing their ability to mitigate possible hazards. Because text mining tools and technologies can aggregate relevant data from hundreds of text data

sources and build linkages between the retrieved insights, they enable companies to access the appropriate information at the right time, therefore improving the risk management framework.

Web Mining

Web Mining is the process of Data Mining techniques to automatically discover and extract information from Web documents and services. The main purpose of web mining is discovering useful information from the World-Wide Web and its usage patterns.

Web mining can be broadly divided into three different types of techniques of mining: Web Content Mining, Web Structure Mining, and Web Usage Mining. These are explained as following below.



Web Content Mining: Web content mining is the application of extracting useful information from the content of the web documents.

Web Structure Mining: Web structure mining is the application of discovering structure information from the web.

Web Usage Mining: Web usage mining is the application of identifying or discovering interesting usage patterns from large data sets.

Applications of Web Mining

The applications of web mining are wide-ranging and include:

Personalized marketing
E-Commerce
Search Engine optimization
Fraud Detection
Sentiment Analysis
Web Content Analysis
Customer Service
Healthcare

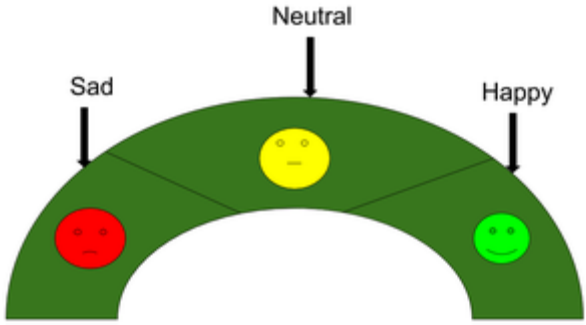
Sentiment Analysis

Sentiment analysis is a popular task in natural language processing. The goal of sentiment analysis is to classify the text based on the mood or mentality expressed in the text, which can be positive negative, or neutral.

Sentiment analysis is the process of classifying whether a block of text is positive, negative, or, neutral. The goal which Sentiment analysis tries to gain is to be analyzed people's opinions in a way that can help businesses expand. It focuses not only on polarity (positive, negative & neutral) but also on emotions (happy, sad, angry, etc.). It uses various Natural Language Processing algorithms such as Rule-based, Automatic, and Hybrid.

Example of Sentiment Analysis

if we want to analyze whether a product is satisfying customer requirements, or is there a need for this product in the market? We can use sentiment analysis to monitor that product's reviews.



Sentiment analysis is also efficient to use when there is a large set of unstructured data, and we want to classify that data by automatically tagging it.Net Promoter Score (NPS) surveys are used extensively to gain knowledge of how a customer perceives a product or service. Sentiment analysis also gained popularity due to its feature to process large volumes of NPS responses and obtain consistent results quickly.

Types of Sentiment Analysis

Fine-grained sentiment analysis: This depends on the polarity base. This category can be designed as very positive, positive, neutral, negative, or very negative. The rating is done on a scale of 1 to 5. If the rating is 5 then it is very positive, 2 then negative, and 3 then neutral.

Emotion detection: The sentiments happy, sad, angry, upset, jolly, pleasant, and so on come under emotion detection. It is also known as a lexicon method of sentiment analysis.

Aspect-based sentiment analysis: It focuses on a particular aspect for instance if a person wants to check the feature of the cell phone then it checks the aspect such as the battery, screen, and camera quality then aspect based is used.

Multilingual sentiment analysis: Multilingual consists of different languages where the classification needs to be done as positive, negative, and neutral. This is highly challenging and comparatively difficult.

Applications

Sentiment Analysis has a wide range of applications as:

Social Media: If for instance the comments on social media side as Instagram, over here all the reviews are analyzed and categorized as positive, negative, and neutral.

Customer Service: In the play store, all the comments in the form of 1 to 5 are done with the help of sentiment analysis approaches.

Marketing Sector: In the marketing area where a particular product needs to be reviewed as good or bad.

Reviewer side: All the reviewers will have a look at the comments and will check and give the overall review of the product.

Social Media Analytics

Social Media has been around for 30 years, but the rise in the user base is recent. The data from social media platforms can be used to boost business. This is where Social Media Mining takes over. The amount of data these companies have to deal

with is huge and scattered. Social Media Mining helps in extracting meaning from the big data.

Social media mining can help us get insights to study customer behavior and interests, systematize and using this information, you can serve better and compound your earnings.

Benefits of Social Media Mining

1. Spot trends before they become trend
2. Sentiment Analysis
3. Keyword Identification
4. Create a better product
5. Competitor Analysis
6. Event Identification
7. Managing Real-time Events
8. Provide Useful Content and Stop Spamming
9. Recognize behaviour