UNIT 3

INTRODUCTION TO DATA MINING

Data Mining: Concepts and applications - Data mining process - Text & Web Analytics: Text analytics and text mining overview- Text mining applications- Web mining overview- Social media analytics- Sentiment analysis overview- Big Data Analytics: Definition and characteristics of big data- Fundamentals of big data analytics.

3.1 Data Mining: Concepts and applications

Data mining is an important role for IT professionals, and a degree in data analytics can help you be qualified to have a career in data mining. But everyone in business also needs to understand data mining—it is vital to how many business processes are done and how information is gleaned, so current and aspiring business professionals need to understand how this process works as well.

What is data mining?

Simply put, data mining is the process that companies use to turn raw data into useful information. They utilize software to look for patterns in large batches of data so they can learn more about customers. It pulls out information from data sets and compares it to help the business make decisions. This eventually helps them to develop strategies, increase sales, market effectively, and more.

Data mining sometimes gets confused with machine learning and data analysis, but these terms are all very different and unique.

While both data mining and machine learning use patterns and analytics, data mining looks for patterns that already exist in data, while machine learning goes beyond to predict future outcomes based on the data. In data mining, the “rules” or patterns aren’t known from the start. In many cases of machine learning, the machine is given a rule or variable to understand the data. Additionally data mining relies on human intervention and decisions, but machine learning is meant to be started by a human and then learn on its own. There is quite a bit of overlap between data mining and machine learning, machine learning processes are often utilized in data mining in order to automate those processes.

Similarly data analysis and data mining aren’t interchangeable terms. Data mining is used in data analytics, but they aren’t the same. Data mining is the process of getting the information from large data sets, and data analytics is when companies take this information and dive into it to learn more. Data analysis involves inspecting, cleaning, transforming, and modeling data. The ultimate goal of analysis is discovering useful information, informing conclusions, and making decisions.

Data mining, data analysis, artificial intelligence, machine learning, and many other terms are all combined in business intelligence processes that help a company or organization make decisions and learn more about their customers and potential outcomes.

Overview of the data mining process.

Almost all businesses use data mining, and it's important to understand the data mining process and how it can help a business make decisions.

Business understanding. The first step to successful data mining is to understand the overall objectives of the business, then be able to convert this into a data mining problem and a plan. Without an understanding of the ultimate goal of the business, you won't be able to design a good data mining algorithm. For example, a supermarket may want to use data mining to learn more about their customers. The business understanding is that a supermarket is looking to find out what their customers are buying the most.

Data understanding. After you know what the business is looking for, it's time to collect data. There are many complex ways that data can be obtained from an organization, organized, stored, and managed. Data mining involves getting familiar with the data, identifying any issues, getting insights, or observing subsets. For example, the supermarket may use a rewards program where customers can input their phone number when they purchase, giving the supermarket access to their shopping data.

Data Preparation. Data preparation involves getting the information production ready. This is the biggest part of data mining. It is taking the computer-language data, and converting it into a form that people can understand and quantify. Transforming and cleaning the data for modeling is key for this step.

Modeling. In the modeling phase, mathematical models are used to search for patterns in the data. There are usually several techniques that can be used for the same set of data. There is a lot of trial and error involved in modeling.

Evaluation. When the model is complete, it needs to be carefully evaluated and the steps to make the model need to be reviewed, to ensure it meets the business objectives. At the end of this phase, a decision about the data mining results will be made. In the supermarket example, the data mining results will provide a list of what the customer has purchased, which is what the business was looking for.

Deployment. This can be a simple or complex part of data mining, depending on the output of the process. It can be as simple as generating a report, or as complex as creating a repeatable data mining process to happen regularly.

After the data mining process has been completed, a business will be able to make their decisions and implement changes based on what they have learned.

10 Key Data Mining Techniques and How Businesses Use Them

Businesses collect and store an unimaginable amount of data, but how do they turn all that data into insights that help them build a better business? Data mining, the process of sifting through massive

amounts of data to identify hidden business trends or patterns, makes these transformational business insights possible.
Data mining is not a new technology. The advent of modern computers and application of data mining techniques meant businesses could finally analyze exponential amounts of data and extract non-intuitive, valuable insights; forecasting likely business outcomes, mitigating risks, and taking advantage of newly identified opportunities.

So what are the key techniques that aspiring data miners should know? Here are 10 data mining techniques that we will explore in detail:

1. Clustering
2. Association
3. Data Cleaning
4. Data Visualization
5. Classification
6. Machine Learning
7. Prediction
8. Neural Networks
9. Outlier Detection
10. Data Warehousing

. Clustering

Clustering is a technique used to represent data visually — such as in graphs that show buying trends or sales demographics for a particular product.

What Is Clustering in Data Mining?

Clustering refers to the process of grouping a series of different data points based on their characteristics. By doing so, data miners can seamlessly divide the data into subsets, allowing for more informed decisions in terms of broad demographics (such as consumers or users) and their respective behaviors.

Methods for Data Clustering

- **Partitioning method:** This involves dividing a data set into a group of specific clusters for evaluation based on the criteria of each individual cluster. In this method, data points belong to just one group or cluster.
- **Hierarchical method:** With the hierarchical method, data points are a single cluster, which are grouped based on similarities. These newly created clusters can then be analyzed separately from each other.
- **Density-based method:** A machine learning method where data points plotted together are further analyzed, but data points by themselves are labeled “noise” and discarded.
- **Grid-based method:** This involves dividing data into cells on a grid, which then can be clustered by individual cells rather than by the entire database. As a result, grid-based clustering hase a fast processing time.
- **Model-based method:** In this method, models are created for each data cluster to locate the best data to fit that particular model.

Examples of Clustering in Business

Clustering helps businesses manage their data more effectively. For example, retailers can use clustering models to determine which customers buy particular products, on which days, and with what frequency. This can help retailers target products and services to customers in a specific demographic or region. Clustering can help grocery stores group products by a variety of characteristics (brand, size, cost, flavor, etc.) and better understand their sales tendencies. It can also help car insurance companies that want to identify a set of customers who typically have high annual claims in order to price policies more effectively. In addition, banks and financial institutions might use clustering to better understand how customers use in-person versus virtual services to better plan branch hours and staffing.

2. Association

Association rules are used to find correlations, or associations, between points in a data set.

What Is Association in Data Mining?

Data miners use association to discover unique or interesting relationships between variables in databases. Association is often employed to help companies determine marketing research and strategy.

Methods for Data Mining Association

Two primary approaches using association in data mining are the single-dimensional and multi-dimensional methods.

- **Single-dimensional association:** This involves looking for one repeating instance of a data point or attribute. For instance, a retailer might search its database for the instances a particular product was purchased.
- **Multi-dimensional association:** This involves looking for more than one data point in a data set. That same retailer might want to know more information than what a customer purchased — such as their age, method of purchase (cash or credit card), or age.

Examples of Association in Business:

The analysis of impromptu shopping behavior is an example of association — that is, retailers notice in data studies that parents shopping for childcare supplies are more likely to purchase specialty food or beverage items for themselves during the same trip. These purchases can be analyzed through statistical association.

Association analysis carries many other uses in business. For retailers, it's particularly helpful in making purchasing suggestions. For example, if a customer buys a smartphone, tablet, or video game device, association analysis can recommend related items like cables, applicable software, and protective cases. Additionally, association is used by the government to employ census data and plan for public services; it is also used by doctors to diagnose various illnesses and conditions more effectively.

3. Data Cleaning

Data cleaning is the process of preparing data to be mined.

What Is Data Cleaning in Data Mining?

Data cleaning involves organizing data, eliminating duplicate or corrupted data, and filling in any null values. When this process is complete, the most useful information can be harvested for analysis.

Methods for Data Cleaning

- **Verifying the data:** This involves checking that each data point in the data set is in the proper format (e.g, telephone numbers, social security numbers).
- **Converting data types**: This ensures data is uniform across the data set. For instance, numeric variables only contain numbers, while string variables can contain letters, numbers, and characters.
- **Removing irrelevant data:** This clears useless or inapplicable data so full emphasis can be placed on necessary data points.
- **Eliminating duplicate data points:** This helps speed up the mining process by boosting efficiency and reducing errors.
- **Removing errors:** This eliminates typing mistakes, spelling errors, and input errors that could negatively affect analysis outcomes.
- **Completing missing values:** This provides an estimated value for all data and reduces missing values, which can lead to skewed or incorrect results.

Examples of Data Cleaning in Business

According to Experian, 95 percent of businesses say they have been impacted by poor data quality. Working with incorrect data wastes time and resources, increases analysis costs (because models need to be repeated), and often leads to faulty analytics.

Ultimately, no matter how great their models or algorithms are, businesses suffer when their data is incorrect, incomplete, or corrupted.

4. Data Visualization

Data visualization is the translation of data into graphic form to illustrate its meaning to business stakeholders.

What Is Data Visualization in Data Mining?

Data can be presented in visual ways through charts, graphs, maps, diagrams, and more. This is a primary way in which data scientists display their findings.

Methods for Data Visualization

Many methods exist for representing data visually. Here are a few:

- **Comparison charts:** Charts and tables express relationships in the data, such as monthly product sales over a one-year period.
- **Maps:** Data maps are used to visualize data pertaining to specific geographic locations. Through maps, data can be used to show population density and changes; compare populations of neighboring states, counties, and countries; detect how populations are spread over geographic regions; and compare characteristics in one region to those in other regions.
- **Heat maps:** This is a popular visualization technique that represents data through different colors and shading to indicate patterns and ranges in the data. It can be used to track everything from a region's temperature changes to its food and pop culture trends.
- **Density plots:** These visualizations track data over a period of time, creating what can look like a mountain range. Density plots make it easy to represent occurrences of single events over time (e.g., month, year, decade).
- **Histograms:** These are similar to density plots but are represented by bars on a graph instead of a linear form.

- **Network diagrams:** These diagrams show how data points relate to each other by using a series of lines (or links) to connect objects together.
- **Scatter plots:** These graphs represent data point relationships on a two-variable axis. Scatter plots can be used to compare unique variables such as a country's life expectancy or the amount of money spent on healthcare annually.
- **Word clouds:** These graphics are used to highlight specific word or phrase instances appearing in a body of text; the larger the word's size in the cloud, the more frequent its use.

Examples of Data Visualization in Business

Representing data visually is an important skill because it makes data readily understandable to executives, clients, and customers. According to Markets and Markets, the market size for global data visualization tools is expected to nearly double (to $10.2 billion) by 2026.

Companies can make faster, more informed decisions when presented with data that is easy to understand and interpret. Today, this is typically accomplished through effective, visually accessible mediums such as graphs, 3D models, and even augmented reality. As a result, it's a good idea for aspiring data professionals to consider learning such skills through a data science and visualization bootcamp.

5. Classification

Classification is a fundamental technique in data mining and can be applied to nearly every industry. It is a process in which data points from large data sets are assigned to categories based on how they're being used.

What Is Classification in Data Mining?

In data mining, classification is considered to be a form of clustering — that is, it is useful for extracting comparable points of data for comparative analysis. Classification is also used to designate broad groups within a demographic, target audience, or user base through which businesses can gain stronger insights.

Methods for Data Mining Classification

- **Logistic regression:** This algorithm attempts to show the probability of a specific outcome within two possible results. For example, an email service can use logistic regression to predict whether or not an email is spam.
- **Decision trees:** Once data is classified, follow-up questions can be asked, and the results diagrammed into a chart called a decision tree. For example, if a computer company wants to predict the likelihood of laptop purchases, it may ask, *Is the potential buyer a student?* The data is classified into "Yes" and "No" decision trees, with other questions to be asked afterward in a similar fashion.
- **K-nearest neighbors (KNN):** This is an algorithm that tries to identify an unknown object by comparing it to others. For instance, grocery chains might use the K-nearest neighbors algorithm to decide whether to include a sushi or hot meals station in their new store layout based on consumer habits in the local marketplace.
- **Naive Bayes:** Based on the Bayes Theorem of Probability, this algorithm uses historical data to predict whether similar events will occur based on a different set of data.
- **Support Vector Machine (SVM):** This machine learning algorithm is often used to define the line that best divides a data set into two classes. An SVM can help classify images and is used in facial and handwriting recognition software.

Examples of Classification in Business

Financial institutions classify consumers based on many variables to market new loans or project credit card risks. Meanwhile, weather apps classify data to project snowfall totals and other similar figures. Grocery stores also use classification to group products by the consumers who buy them, helping forecast buying patterns.

6. Machine Learning

Machine learning is the process by which computers use algorithms to learn on their own. An increasingly relevant part of modern technology, machine learning makes computers "smarter" by teaching them how to perform tasks based on the data they have gathered.

What Is Machine Learning in Data Mining?

In data mining, machine learning's applications are vast. Machine learning and data mining fall under the umbrella of data science but aren't interchangeable terms. For instance, computers perform data mining as part of their machine learning functions.

Methods for Machine Learning

- **Supervised learning:** In this method, algorithms train machines to learn using pre-labeled data with correct values, which the machines then classify on their own. It's called supervised because the process trains (or "supervises") computers to classify data and predict outcomes. Supervised machine learning is used in data mining classification.
- **Unsupervised learning:** When computers handle unlabeled data, they engage in unsupervised learning. In this case, the computer classifies the data itself and then looks for patterns on its own. Unsupervised models are used to perform clustering and association.
- **Semi-supervised learning:** Semi-supervised learning uses a combination of labeled and unlabeled data, making it a hybrid of the above models.
- **Reinforcement learning:** This is a more layered process in which computers learn to make decisions based on examining data in a specific environment. For example, a computer might learn to play chess by examining data from thousands of games played online.

Examples of Machine Learning in Business

With machine learning, companies can use computers to quickly identify all sorts of data patterns (in sales, product usage, buying habits, etc.) and develop business plans using those insights. This is a growing need in many industries.

7. Neural Networks

Computers process large amounts of data much faster than human brains but don't yet have the capacity to apply common sense and imagination in working with the data. Neural networks are one way to help computers reason more like humans.

What Are Neural Networks in Data Mining?

Artificial neural networks attempt to digitally mimic the way the human brain operates. Neural networks combine many computer processors (similar to the way the brain uses neurons) to process data, make decisions, and learn as a human would — or at least as closely as possible.

Neural Network Methods

Neural networks consist of three main layers: input, "hidden," and output. Data enters through the input layer, is processed in the hidden layer, and is resolved in the output layer where any relevant action based

on the data is then taken. The hidden layer can consist of many processing layers, depending on the amount of data being used and learning taking place.
Supervised and unsupervised learning also apply to neural networks; neural networks use these types of algorithms to "train" themselves to function in ways similar to the human brain.

Examples of Neural Networks in Business

Neural networks have a wide range of applications. They can help businesses predict consumer buying patterns and focus marketing campaigns on specific demographics. They can also help retailers make accurate sales forecasts and understand how to use dynamic pricing. Furthermore, they help to improve diagnostic and treatment methods in healthcare, improving care and performance.

8. Outlier Detection

Outlier detection is a key component of maintaining safe databases. Companies use it to test for fraudulent transactions, such as abnormal credit card usage that might suggest theft.

What Is Outlier Detection in Data Mining?

While other data mining methods seek to identify patterns and trends, outlier detection looks for the unique: the data point or points that differ from the rest or diverge from the overall sample. Outlier detection finds errors, such as data that was input incorrectly or extracted from the wrong sample. Natural data deviations can be instructive as well.

Methods for Outlier Detection

- **Numeric outlier:** Outliers are detected based on the Interquartile Range, or the middle 50 percent of values. Data points outside that range are considered outliers.
- **Z-score:** The Z-Score denotes how many standard deviations a data point is from the sample's mean. This is also known as extreme value analysis.
- **DBSCAN:** This stands for "density-based spatial clustering of applications with noise" and is a method that defines data as core points, border points, and noise points, which are the outliers.
- **Isolation forest:** This method isolates anomalies in large sets of data (the forest) with an algorithm that searches for those anomalies instead of profiling normal data points.

Examples of Outlier Detection in Business

Almost every business can benefit from understanding anomalies in their production or distribution lines and how to fix them. Retailers can use outlier detection to learn why their stores witness an odd increase in purchases, such as snow shovels being bought in the summer, and how to respond to such findings. Generally, outlier detection is employed to enhance logistics, instill a culture of preemptive damage control, and create a smoother environment for customers, users, and other key groups.

9. Prediction

Predictive modeling seeks to turn data into a projection of future action or behavior. These models examine data sets to find patterns and trends, then calculate the probabilities of a future outcome.

What Is Prediction in Data Mining?

Predictive modeling is among the most common uses of data mining and works best with large data sets that represent a broad sample size.

Methods for Prediction

Predictive modeling uses some of the same techniques and terminology as other data mining processes. Here are four examples:

- **Forecast modeling:** This is a common technique in which the computer answers a question (for instance, *How much milk should a store have in stock on Monday?*) by analyzing historical data.
- **Classification modeling:** Classification places data into groups where it can be used to answer direct questions.
- **Cluster modeling:** By clustering data into groups with shared characteristics, a predictive model can be used to study those data sets and make decisions.
- **Time series modeling:** This model analyzes data based on when the data was input. A study of sales trends over a year is an example of time series modeling.

How does data mining inform business analytics?

So why is data mining important for businesses? Businesses that utilize data mining are able to have a competitive advantage, better understanding of their customers, good oversight of business operations, improved customer acquisition, and new business opportunities. Different industries will have different benefits from their data analytics. Some industries are looking for the best ways to get new customers, others are looking for new marketing techniques, and others are working to improve their systems. The data mining process is what gives businesses the opportunities and understanding for how to make their decisions, analyze their information, and move forward.

Data mining techniques in business analytics.

Now that you understand why data mining is important, it's beneficial to see how data mining works specifically in business settings.

Classification. This data mining technique is more complex, using attributes of data to move them into discernable categories, helping you draw further conclusions. Supermarket data mining may use classification to group the types of groceries customers are buying, like produce, meat, bakery items, etc. These classifications help the store learn even more about customers, outputs, etc.

Clustering. This technique is very similar to classification, chunking data together based on their similarities. Cluster groups are less structured than classification groups, making it a more simple option for data mining. In the supermarket example, a simple cluster group could be food and non-food items instead of the specific classes.

Association rules. Association in data mining is all about tracking patterns, specifically based on linked variables. In the supermarket example, this may mean that many customers who buy a specific item may also buy a second, related item. This is how stores may know how to group certain food items together, or in online shopping they may show "people also bought this" section.

Regression analysis. Regression is used to plan and model, identifying the likelihood of a specific variable. The supermarket may be able to project price points based on availability, consumer demand, and their competition. Regression helps data mining by identifying the relationship between variables in a set.

Anomaly/outlier detection. For many data mining cases, just seeing the overarching pattern might not be all you need. Data needs to be able to identify and understand the outliers in your data as well. For example, in the supermarket if most of the shoppers are female, but one week in February is mostly men, you'll want to investigate that outlier and understand what is behind it.

Free data mining tools for businesses.

DataMelt. DataMelt performs mathematics, statistics, calculations, data analysis, and visualization. Many scripting languages and Java packages are available in this system.

ELKI Data Mining Framework. ELKI focuses on algorithms with a specific emphasis on unsupervised cluster and outlier systems. ELKI is designed to be easy for researchers, students, and business organizations to use

Orange Data Mining. Orange data mining helps organizations do simple data analysis and use top visualization and graphics. Heatmaps, hierarchical clustering, decision trees, and more are used in this process.

The R Project for Statistical Computing. The R Project is used in statistical modeling and graphics and is utilized on many operating systems and programs

Rattle GUI. Rattle GUI presents statistical and visual summaries of data, helps prepare it to be modeled, and utilizes supervised and unsupervised machine learning to present the information.

3.2 Data mining process

Defining the problem

Data mining processes start with a clearly defined business problem. For instance, increase sales or get more return customers can be business problems to study.

Selecting features/variables

For instance, businesses collect data based on the customer and what they have purchased from the company to examine returning customers and create customer profiles. Age, location, and income would be helpful variables to include in the selected dataset for curation.

Collecting and curating data

Once the question and dataset are determined, data engineers create the data pipeline to collect the data or put existing data in the desired format. Depending on the problem, the dataset is curated to give insight about this specific business problem.

Analyzing the data

Data scientists investigate the data to remove outliers or anomalies and analyze it to determine patterns in order to help solve the business problem.

Make business decisions and changes

Once the results are out, the BA team can makes data-driven decisions to change or optimize a certain business strategy or operation.

Track changes

Based on the decision, the data collection and analysis processes continue to understand if the decisions work as expected.

Adjust and repeat

If the results are as expected, BA teams can generalize the changes to optimize similar processes or strategies. If the results are not desirable, BA teams can do further testing to understand why changes did not work and adjust the strategies.

Some of the benefits of applying data mining for business analytics include:

Customer benefits:

Understanding the customer landscape

Increasing marketing effectiveness

Enhancing customer experience

Operational benefits:

Improving Operational Efficiency

Making data-driven decisions

Retaining employees

Business benefits:

Examining the competition

Expanding sales and increasing Revenue

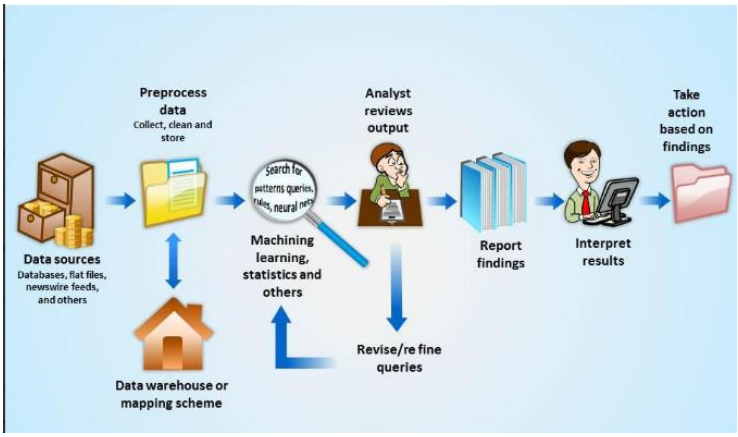

Fig 3.2 Data Mining Process

3.3 Text & Web Analytics: Text analytics and text mining overview

Researchers in this stream specialise in the use of data mining and statistical machine learning to analyse structured and unstructured data. We have specific expertise in text mining, social media mining, social network analysis, predictive analytics, smart business and recommender systems.

Our text analytics work is focused on extracting information from unstructured text to create structured data patterns. Our web analytics research is focused on collecting, analysing and reporting web data for the purpose of understanding and optimising web usage. This work provides new and exciting business insights into customer and online activities.

Text Analytics

Entity extraction, text categorization and text clustering

Document summarisation

Topic model and latent semantic analysis

Topic discover and public event detection

Search, retrieval and ranking

Microblog and twitter mining

Short text analysis and semantic enhancement

Opinion mining and sentiment analysis

Social spammer detection and social influence analysis

Web analytics

Customer behaviour and access pattern mining

Customer profiling and segmentation

Customer retention and churn analysis

Sales trend analysis and sales forecasting

Marketing segmentation and cross-sale strategies

Link analysis and link prediction

User community detection and evolution

Spatial-temporal analysis

Recommender systems

Content-based, collaborative filtering, matrix factorisation algorithms

Social recommender systems

Cross-domain recommendation

Location-based social networks, and point-of-interest recommendation

Group-based recommendation

Contextual-aware recommendation

Mobile and handheld device-based recommendation systems

Fig 3.3 text mining and text analytics – overview

Benefits of Text Analytics

There are a range of ways that text analytics can help businesses, organizations, and event social movements:

Help businesses to understand customer trends, product performance, and service quality. This results in quick decision making, enhancing business intelligence, increased productivity, and cost savings.

Helps researchers to explore a great deal of pre-existing literature in a short time, extracting what is relevant to their study. This helps in quicker scientific breakthroughs.

Assists in understanding general trends and opinions in the society that enable governments and political bodies in decision making.

Text analytic techniques help search engines and information retrieval systems to improve their performance, thereby providing fast user experiences.

Refine user content recommendation systems by categorizing related content.

Text Analytics Techniques and Use Cases

There are several techniques related to analyzing the unstructured text. Each of these techniques is used for different use case scenarios.

Sentiment analysis

Sentiment analysis is used to identify the emotions conveyed by the unstructured text. The input text includes product reviews, customer interactions, social media posts, forum discussions, or blogs. There are different types of sentiment analysis. Polarity analysis is used to identify if the text expresses positive or negative sentiment. The categorization technique is used for a more fine-grained analysis of emotions - confused, disappointed, or angry.

Use cases of sentiment analysis:

Measure customer response to a product or a service

Understand audience trends towards a brand

Understand new trends in consumer space

Prioritize customer service issues based on the severity

Track how customer sentiment evolves over time

Topic modelling

This technique is used to find the major themes or topics in a massive volume of text or a set of documents. Topic modeling identifies the keywords used in text to identify the subject of the article.

Use cases of topic modeling:

Large law firms use topic modeling to examine hundreds of documents during large litigations.

Online media uses topic modeling to pick up trending topics across the web.

Researchers use topic modeling for exploratory literature review.

Businesses can determine which of their products are successful.

Topic modeling helps anthropologists to determine the emergent issues and trends in a society based on the content people share on the web.

Named Entity Recognition (NER)

NER is a text analytics technique used for identifying named entities like people, places, organizations, and events in unstructured text. NER extracts nouns from the text and determines the values of these nouns.

Use cases of named entity recognition:

NER is used to classify news content based on people, places, and organizations featured in them.

Search and recommendation engines use NER for information retrieval.

For large chain companies, NER is used to sort customer service requests and assign them to a specific city, or outlet.

Hospitals can use NER to automate the analysis of lab reports.

Term frequency – inverse document frequency

TF-IDF is used to determine how often a term appears in a large text or group of documents and therefore that term's importance to the document. This technique uses an inverse document frequency factor to filter out frequently occurring yet non-insightful words, articles, propositions, and conjunctions.

Event extraction

This is a text analytics technique that is an advancement over the named entity extraction. Event extraction recognizes events mentioned in text content, for example, mergers, acquisitions, political moves, or important meetings. Event extraction requires an advanced understanding of the semantics of text content. Advanced algorithms strive to recognize not only events but the venue, participants, date, and time wherever applicable. Event extraction is a beneficial technique that has multiple uses across fields.

Use cases of event extraction:

Link analysis: This is a technique to understand "who met whom and when" through event extraction from communication over social media. This is used by law enforcement agencies to predict possible threats to national security.

Geospatial analysis: When events are extracted along with their locations, the insights can be used to overlay them on a map. This is helpful in the geospatial analysis of the events.

Business risk monitoring: Large organizations deal with multiple partner companies and suppliers. Event extraction techniques allow businesses to monitor the web to find out if any of their partners, like suppliers or vendors, are dealing with adverse events like lawsuits or bankruptcy.

Steps Involved with Text Analytics

Text analytics is a sophisticated technique that involves several pre-steps to gather and cleanse the unstructured text. There are different ways in which text analytics can be performed. This is an example of a model workflow.

Data gathering - Text data is often scattered around the internal databases of an organization, including in customer chats, emails, product reviews, service tickets and Net Promoter Score surveys. Users also generate external data in the form of blog posts, news, reviews, social media posts and web forum discussions. While the internal data is readily available for analytics, the external data needs to be gathered.

Preparation of data - Once the unstructured text data is available, it needs to go through several preparatory steps before machine learning algorithms can analyze it. In most of the text analytics software, this step happens automatically. Text preparation includes several techniques using natural language processing as follows:

Tokenization: In this step, the text analysis algorithms break the continuous string of text data into tokens or smaller units that make up entire words or phrases. For instance, character tokens could be each individual letter in this word: F-I-S-H. Or, you can break up by subword tokens: Fish-ing. Tokens represent the basis of all natural language processing. This step also discards all the unwanted contents of the text, including white spaces.

Part-of-speech-tagging: In this step, each token in the data is assigned a grammatical category like noun, verb, adjective, and adverb.

Parsing: Parsing is the process of understanding the syntactical structure of the text. Dependency parsing and constituency parsing are two popular techniques used to derive syntactical structure.

Lemmatization and stemming: These are two processes used in data preparation to remove the suffixes and affixes associated with the tokens and retain its dictionary form or lemma.

Stopword removal: This is the phase when all the tokens that have frequent occurrence but bear no value in the text analytics. This includes words such as ‘and’, ‘the’ and ‘a’.

Text analytics - After the preparation of unstructured text data, text analytics techniques can now be performed to derive insights. There are several techniques used for text analytics. Prominent among them are text classification and text extraction.

Text classification: This technique is also known as text categorization or tagging. In this step, certain tags are assigned to the text based on its meaning. For example, while analyzing customer reviews, tags like “positive” or “negative” are assigned. Text classification often is done using rule-based systems or machine learning-based systems. In rule-based systems, humans define the association between language pattern and a tag. “Good” may indicate positive review; “bad” may idenitfy a negative review.

Machine learning systems use past examples or training data to assign tags to a new set of data. The training data and its volume are crucial, as larger sets of data helps the machine learning algorithms to give accurate tagging results. The main algorithms used in text classification are Support Vector Machines (SVM), Naive Bayes family of algorithms (NB), and deep learning algorithms.

Text extraction: This is the process of extracting recognizable and structured information from the unstructured input text. This information includes keywords, names of people, places and events. One of the simple methods for text extraction is regular expressions. However, this is a complicated method to maintain when the complexity of input data increases. Conditional Random Fields (CRF) is a statistical method used in text extraction. CRF is a sophisticated but effective way of extracting vital information from the unstructured text.

3.4 Text mining applications

Sentiment analysis is a widely used text mining application that can track customer sentiment about a company. Also known as opinion mining, sentiment analysis mines text from online reviews, social networks, emails, call center interactions and other data sources to identify common threads that point to positive or negative feelings on the part of customers. Such information can be used to fix product issues, improve customer service and plan new marketing campaigns, among other things.

Other common text mining uses include screening job candidates based on the wording in their resumes, blocking spam emails, classifying website content, flagging insurance claims that may be fraudulent, analyzing descriptions of medical symptoms to aid in diagnoses, and examining corporate documents as part of electronic discovery processes. Text mining software also offers information retrieval capabilities akin to what search engines and Enterprise search platforms provide, but that's usually just an element of higher level text mining applications, and not a use in and of itself.

Chatbots answer questions about products and handle basic customer service tasks; they do so by using natural language understanding (NLU) technology, a subcategory of NLP that helps the bots understand human speech and written text so they can respond appropriately.

Natural language generation (NLG) is another related technology that mines documents, images and other data, and then creates text on its own. For example, NLG algorithms are used to write descriptions of neighborhoods for real estate listings and explanations of key performance indicators tracked by business intelligence systems.

Benefits of text mining

Using text mining and analytics to gain insight into customer sentiment can help companies detect product and business problems and then address them before they become big issues that affect sales. Mining the text in customer reviews and communications can also identify desired new features to help strengthen product offerings. In each case, the technology provides an opportunity to improve the overall customer experience, which will hopefully result in increased revenue and profits.

Text mining can also help predict customer churn, enabling companies to take action to head off potential defections to business rivals as part of their marketing and customer relationship management programs. Fraud detection, risk management, online advertising and web content management are other functions that can benefit from the use of text mining tools.

In healthcare, the technology may be able to help diagnose illnesses and medical conditions in patients based on the symptoms they report.

Text mining challenges and issues

Text mining can be challenging because the data is often vague, inconsistent and contradictory. Efforts to analyze it are further complicated by ambiguities that result from differences in syntax and semantics, as well as the use of slang, sarcasm, regional dialects and technical language specific to individual vertical industries. As a result, text mining algorithms must be trained to parse such ambiguities and inconsistencies when they categorize, tag and summarize sets of text data.

In addition, the deep learning models used in many text mining applications require large amounts of training data and processing power, which can make them expensive to run. Inherent bias in data sets is another issue that can lead deep learning tools to produce flawed results if data scientists don't recognize the biases during the model development process.

There's also a lot of text mining software to choose from. Dozens of commercial and open source technologies are available, including tools from major software vendors, including IBM, Oracle, SAS, SAP and Tibco.

3.8 Big Data Analytics: Definition and characteristics of big data

Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes from terabytes to zetta bytes.

Big data is a term applied to data sets whose size or type is beyond the ability of traditional relational databases to capture, manage and process the data with low latency. Big data has one or more of the following characteristics: high volume, high velocity or high variety. Artificial intelligence (AI), mobile, social and the Internet of Things (IoT) are driving data complexity through new forms and sources of data. For example, big data comes from sensors, devices, video/audio, networks, log files, transactional applications, web, and social media much of it generated in real time and at a very large scale.

Analysis of big data allows analysts, researchers and business users to make better and faster decisions using data that was previously inaccessible or unusable. Businesses can use advanced analytics techniques such as text analytics, machine learning, predictive analytics, data mining, statistics and natural language processing to gain new insights from previously untapped data sources independently or together with existing enterprise data.

The Lifecycle Phases of Big Data Analytics now, let's review how Big Data analytics works:

Stage 1 - Business case evaluation - The Big Data analytics lifecycle begins with a business case, which defines the reason and goal behind the analysis.

Stage 2 - Identification of data - Here, a broad variety of data sources are identified.

Stage 3 - Data filtering - All of the identified data from the previous stage is filtered here to remove corrupt data.

Stage 4 - Data extraction - Data that is not compatible with the tool is extracted and then transformed into a compatible form.

Stage 5 - Data aggregation - In this stage, data with the same fields across different datasets are integrated.

Stage 6 - Data analysis - Data is evaluated using analytical and statistical tools to discover useful information.

Stage 7 - Visualization of data - With tools like Tableau, Power BI, and QlikView, Big Data analysts can produce graphic visualizations of the analysis.

Stage 8 - Final analysis result - This is the last step of the Big Data analytics lifecycle, where the final results of the analysis are made available to business stakeholders who will take action.

Benefits and Advantages of Big Data Analytics

1. Risk Management

Use Case: Banco de Oro, a Phillippine banking company, uses Big Data analytics to identify fraudulent activities and discrepancies. The organization leverages it to narrow down a list of suspects or root causes of problems.

2. Product Development and Innovations

Use Case: Rolls-Royce, one of the largest manufacturers of jet engines for airlines and armed forces across the globe, uses Big Data analytics to analyze how efficient the engine designs are and if there is any need for improvements.

3. Quicker and Better Decision Making Within Organizations

Use Case: Starbucks uses Big Data analytics to make strategic decisions. For example, the company leverages it to decide if a particular location would be suitable for a new outlet or not. They will analyze several different factors, such as population, demographics, accessibility of the location, and more.

4. Improve Customer Experience

Use Case: Delta Air Lines uses Big Data analysis to improve customer experiences. They monitor tweets to find out their customers' experience regarding their journeys, delays, and so on. The airline identifies negative tweets and does what's necessary to remedy the situation. By publicly addressing these issues and offering solutions, it helps the airline build good customer relations

3.9 Fundamentals of big data analytics:

Fundamentals of Big Data Analytics are:

It is an essential revolution in the sector of IT, and this technique is enlarging every year. It is the process of inspecting the huge data sets to emphasize both the patterns and insights.

Cost Reduction: The analytics technique like a Cloud Computing, Hadoop which it is important to cost benefits storing into the huge sets of information and data. In reality, they will recognize efficient ways of running the business.

Faster, Best Decision Making: Speed of Hadoop, network classes and that combination of able to analyzing the latest sources of information, business.

Services and Products: The ability to measure client satisfaction and needs through an analytics. They are so many companies are developing the new services and products to meet their client needs.

Real-Time Benefits in Fundamentals of Big Data Analytics

It has been massive growth in this sector, and it led to the usability of big data in numerous industries ranging. For the purpose of, this tool helps Apache Hadoop to minimize the cost of storage.

Banking

Healthcare

Energy

Technology

Consumer

Manufacturing

Most of the banking sectors are using this big data technique via data analytics course. To put it another way, the education field may apply the big data concepts. As well as, a possibility for both the analysis and research utilize the data.

Advantages of Big Data Analytics

Data procurement: Particularly, it is a large amount of data for developing the store. They are several websites are accumulating into the data, secondary, and primary.

Data Integration and Data Quality: The data and information may store in the high changes in data sets. The big data analytics are a lot of repetition which it is creates the expenses and confusion.

Data Segmentation: It may use to distribute the data in various parameters for example location, age, gender, budget, product segmentation and so on.

Business Intelligence: Especially, Fundamentals of Big Data Analytics is driven which it is consist the decision making, and it enables the scientists to visual data, aggregate, generate helping into the management decisions.

Prescriptive and Predictive Analytics: It allows the various possible activity towards the solutions. In general, the mixture of historical data are found into the CRM, POS, ERP and HR systems may identify the patterns. Applying the algorithms and statistical models capturing the different datasets.

It's widely accepted today that the phrase "big data" implies more than just storing more data. It also means doing more with data. There are arguably too many terms that we use to describe the techniques for "doing more," although big data analytics or data science probably come closest. We can probably refine the various techniques into three big groups:

Predictive analytics, which are the class of algorithms that use data from the past to predict the future

Collective intelligence, which uses the inputs from large groups to create seemingly intelligent behavior

Machine learning, in which programs "learn from experience" and refine their algorithms-based on new information

These are clearly intersecting techniques—collective intelligence often is predictive, while predictive and collective techniques both involve machine learning.

Predictive algorithms take many forms, but a large proportion build on fundamental mathematical concepts taught in high school. Creating a "line of best fit" between two variables involves a fairly simple computation known as linear regression. Once created, the regression formula can be used to predict the value of one variable based on the other. Regression analysis can be extended to more than two variables (multivariate regression), curves (nonlinear regression), categorical predictions (logistic regression), and adjusted to understand seasonal variation (time series analysis).

Collective intelligence is often predictive, while predictive and collective techniques both involve machine learning.

Collective intelligence sounds like a complex academic pursuit, but it's actually something we encounter every day. When Google or another search engine corrects or predicts your searches, it is using the data collected from the billions of other peoples' searches that came before yours.

Machine learning as a general technique includes most of the algorithms employed by predictive and collective solutions. Whenever a system can adjust its behavior based on new input data, it can be said to have learned.

A supervised machine learning algorithm is one that requires some training in order to build a model. For instance, in the case of spam classification algorithms, human beings are generally required to provide examples of spam and non-spam emails. The spam detector uses these examples—called the training set—to create algorithms that can be used to distinguish spam from non-spam. The final test of the algorithm is to provide it with some fresh data—a validation set—to see how well it does.

Unsupervised machine learning requires no training sets, and clustering algorithms fall into this category. A good example is the familiar basket analysis algorithm—if you order three of the four ingredients in a Waldorf salad from Walmart online, the missing ingredient likely will be recommended to you. This is not because Walmart is comparing your order to a recipe book, but because a clustering algorithm has noticed that these four items usually appear together.

Under the hood, there are dozens of algorithms that can be used to perform machine learning. Classification includes techniques such as logistic regression, naive Bayesian analysis, decision trees, K-nearest neighbors, and Support Vector Machines. Clustering algorithms include K-means and hierarchical clustering. Because of the very large number of complicated algorithms —and those that just sound complicated—it is hard for even the most experienced data scientist to pick the correct technique for the data at hand. For that reason, ensemble techniques often are employed to run multiple algorithms on the data and select the resulting model with the best outcomes.

Fig 3.9 Working of Big Data Analytics

The volume of data that one has to deal has exploded to unimaginable levels in the past decade, and at the same time, the price of data storage has systematically reduced. Private companies and research institutions capture terabytes of data about their users' interactions, business, social media, and also sensors from devices such as mobile phones and automobiles. The challenge of this era is to make sense of this sea of data. This is where big data analytics comes into picture.

Big Data Analytics largely involves collecting data from different sources, munge it in a way that it becomes available to be consumed by analysts and finally deliver data products useful to the organization business.

The process of converting large amounts of unstructured raw data, retrieved from different sources to a data product useful for organizations forms the core of Big Data Analytics.

PART A:

1. How does data mining inform business analytics?
2. What are Benefits of text mining?
3. Specify the Use cases of sentiment analysis?
4. Why is social media analytics important?
5. How does data mining inform business analytics?
6. What are applications of data mining analytics?
7. List out the Examples of Outlier Detection in Business?
8. Compare Business Understanding and Data Understanding.
9. List out the Lifecycle Phases of Big Data Analytics?
10. What are the advantages of big data analytics?

PART B:

1. Explain Data mining techniques in business analytics.
2. Discuss Benefits and Advantages of Big Data Analytics in detail.
3. Justify the free data mining tools for businesses in detail.
4. Develop an advanced decision support system for an e-news web portal using web mining
5. Describe Text mining challenges and issues.
6. Explain the working of big data analytics with neat diagram.