✅ Comparison Between Text Mining and Data Mining

Aspect	Text Mining	Data Mining
Definition	Process of extracting meaningful information from unstructured text data.	Process of discovering patterns and knowledge from structured datasets.
Data Type	Unstructured or semi-structured (e.g., emails, articles, social media posts).	Structured data (e.g., databases, spreadsheets, data warehouses).
Input Format	Natural language text, documents, XML, JSON, etc.	Tables, numerical data, categorical data.
Techniques Used	- Natural Language Processing (NLP) - Sentiment Analysis - Named Entity Recognition (NER) - Topic Modeling	- Classification & Clustering - Association Rule Mining - Regression Analysis - Decision Trees
Tools & Libraries	- NLTK, spaCy, Gensim - TextBlob, BERT - Word2Vec, TF-IDF	- RapidMiner, Weka - Python (Pandas, Scikit-learn) - R, SQL, Tableau
Preprocessing Steps	- Tokenization - Stopword Removal - Stemming/Lemmatization - Vectorization (TF-IDF, Word2Vec)	- Data Cleaning - Normalization - Handling Missing Values - Feature Selection
Output	Extracted topics, entities, sentiments, or summaries.	Hidden patterns, trends, predictions, and relationships.
Challenges	- Ambiguity in language - Context understanding - Sarcasm/Irony detection	- Handling large datasets - Data quality issues - Overfitting/Underfitting
Applications	- Sentiment Analysis - Chatbots - Document Classification - Spam Filtering	- Market Basket Analysis - Fraud Detection - Customer Segmentation - Predictive Analytics

Comparison of Measures for Feature Selection

Measure	Focus	Strength	Limitation
Gini Index	Distribution of words across classes	Simple and interpretable	May be biased by class imbalance
Information Gain	Reduction in entropy	Reflects global and local word importance	Computationally expensive for large datasets
Mutual Information	Dependency between word and class	Captures strong correlations	Can overvalue rare words
Chi-Square	Lack of independence	Normalized, suitable for multi-class problems	Assumes sufficient sample size

Difference Between Influence and Homophily

Aspect	Influence	Homophily
Definition	Influence is the process where one individual affects the behavior, opinions, or decisions of another.	Homophily is the tendency of similar individuals to connect based on shared attributes.
Direction	Influence flows **directionally** from one person (the influencer) to another.	Homophily is **mutual**; both individuals connect due to shared similarities.
Nature	Active: It involves a change in behavior or opinion.	Passive: It involves forming connections based on pre-existing similarity.
Driving Factor	Influence arises from **persuasion, authority, or expertise** of one individual.	Homophily arises from **inherent similarities** like demographics, interests, or beliefs.

Formation of Links	Links form when one individual influences another to connect or adopt behaviors.	Links form because individuals already share common attributes.
Example	A social media influencer persuades their followers to buy a product or adopt a habit.	Gamers on a social network connect with each other because they share an interest in gaming.
Observable Effect	Can lead to **behavioral or opinion change**.	Leads to **assortativity** in the network (formation of clusters based on similarity).
Social Network Outcome	Can create **information cascades** or viral effects as influence spreads through the network.	Creates **segmentation or clustering** where similar individuals form dense groups.
Timeframe	Influence can be instantaneous or gradual.	Homophily is a gradual process based on accumulating connections.