What is text mining?

Text mining, also known as text data mining, is the process of transforming unstructured text into a structured format to identify meaningful patterns and new insights. You can use text mining to analyze vast collections of textual materials to capture key concepts, trends and hidden relationships.

By applying advanced analytical techniques, such as Naïve Bayes, Support Vector Machines (SVM), and other deep learning algorithms, companies are able to explore and discover hidden relationships within their unstructured data.

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

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

Since roughly 80% of data in the world resides in an unstructured format (link resides outside ibm.com), text mining is an extremely valuable practice within organizations. Text mining tools and natural language processing (NLP) techniques, like information extraction (link resides outside ibm.com), allow us to transform unstructured documents into a structured format to enable analysis and the generation of high-quality insights. This, in turn, improves the decision-making of organizations, leading to better business outcomes.

Text mining vs. text analytics

The terms, text mining and text analytics, are largely synonymous in meaning in conversation, but they can have a more nuanced meaning. Text mining and text analysis identifies textual patterns and trends within unstructured data through the use of machine learning, statistics, and linguistics. By transforming the data into a more structured format through text mining and text analysis, more quantitative insights can be found through text analytics. Data visualization techniques can then be harnessed to communicate findings to wider audiences.

How does text mining work?

Understanding the text-mining workflow is vital to unlocking the full potential of the methodology. Here, we'll lay out the text-mining process, highlighting each step and its significance to the overall outcome.

Step 1. Information retrieval

The first step in the text-mining workflow is information retrieval, which requires data scientists to gather relevant textual data from various sources (e.g., websites, social media platforms, customer surveys, online reviews, emails and/or internal databases). The data collection process should be tailored to the specific objectives of the analysis. In the case of social media text mining, that means a focus on comments, posts, ads, audio transcripts, etc.

Step 2. Data preprocessing

Once you collect the necessary data, you'll preprocess it in preparation for analysis. Preprocessing will include several sub-steps, including the following:

- Text cleaning: Text cleaning is the process of removing irrelevant characters, punctuation, special symbols and numbers from the dataset. It also includes converting the text to lowercase to ensure consistency in the analysis stage. This process is especially important when mining social media posts and comments, which are often full of symbols, emojis and unconventional capitalization patterns.

- Tokenization: Tokenization breaks down the text into individual units (i.e., words and/or phrases) known as tokens. This step provides the basic building blocks for subsequent analysis.

- Stop-words removal: Stop words are common words that don't have significant meaning in a phrase or sentence (e.g., "the," "is," "and," etc.). Removing stop words helps reduce noise in the data and improve accuracy in the analysis stage.

- Stemming and lemmatization: Stemming and lemmatization techniques normalize words to their root form. Stemming reduces words to their base form by removing prefixes or suffixes, while lemmatization maps words to their dictionary form. These techniques help consolidate word variations, reduce redundancy and limit the size of indexing files.

- Part-of-speech (POS) tagging: POS tagging facilitates semantic analysis by assigning grammatical tags to words (e.g., noun, verb, adjective, etc.), which is particularly useful for sentiment analysis and entity recognition.

- Syntax parsing: Parsing involves analyzing the structure of sentences and phrases to determine the role of different words in the text. For instance, a parsing model could identify the subject, verb and object of a complete sentence.

Step 3. Text representation

In this stage, you'll assign the data numerical values so it can be processed by machine learning (ML) algorithms, which will create a predictive model from the training inputs. These are two common methods for text representation:

- Bag-of-words (BoW): BoW represents text as a collection of unique words in a text document. Each word becomes a feature, and the frequency of occurrence represents its value. BoW doesn't account for word order, instead focusing exclusively on word presence.

- Term frequency-inverse document frequency (TF-IDF): TF-IDF calculates the importance of each word in a document based on its frequency or rarity across the entire dataset. It weighs down frequently occurring words and emphasizes rarer, more informative terms.

Step 4. Data extraction

Once you've assigned numerical values, you will apply one or more text-mining techniques to the structured data to extract insights from social media data. Some common techniques include the following:

- Sentiment analysis: Sentiment analysis categorizes data based on the nature of the opinions expressed in social media content (e.g., positive, negative or neutral). It can be useful for understanding customer opinions and brand perception, and for detecting sentiment trends.

- Topic modeling: Topic modeling aims to discover underlying themes and/or topics in a collection of documents. It can help identify trends, extract key concepts and predict customer interests. Popular algorithms for topic modeling include Latent Dirichlet Allocation (LDA) and non-negative matrix factorization (NMF).

- Named entity recognition (NER): NER extracts relevant information from unstructured data by identifying and classifying named entities (like person names, organizations, locations and dates) within the text. It also automates tasks like information extraction and content categorization.

- Text classification: Useful for tasks like sentiment classification, spam filtering and topic classification, text classification involves categorizing documents into predefined classes or categories. Machine learning algorithms like Naïve Bayes and support vector machines (SVM), and deep learning models like convolutional neural networks (CNN) are frequently used for text classification.

- Association rule mining: Association rule mining can discover relationships and patterns between words and phrases in social media data, uncovering associations that may not be obvious at first glance. This approach helps identify hidden connections and co-occurrence patterns that can drive business decision-making in later stages.

Step 5. Data analysis and interpretation

The next step is to examine the extracted patterns, trends and insights to develop meaningful conclusions. Data visualization techniques like word clouds, bar charts and network graphs can help you present the findings in a concise, visually appealing way.

Step 6. Validation and iteration

It's essential to make sure your mining results are accurate and reliable, so in the penultimate stage, you should validate the results. Evaluate the performance of the text-mining models using relevant evaluation metrics and compare your outcomes with ground truth and/or expert judgment. If necessary, make adjustments to the preprocessing, representation and/or modeling steps to improve the results. You may need to iterate this process until the results are satisfactory.

Step 7. Insights and decision-making

The final step of the text-mining workflow is transforming the derived insights into actionable strategies that will help your business optimize social media data and usage. The extracted knowledge can guide processes like product improvements, marketing campaigns, customer support enhancements and risk mitigation strategies—all from social media content that already exists.

Text mining techniques

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

Information retrieval

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

- **Tokenization:** This is the process of breaking out long-form text into sentences and words called “tokens”. These are, then, used in the models, like bag-of-words, for text clustering and document matching tasks.
- **Stemming:** This refers to the process of separating the prefixes and suffixes from words to derive the root word form and meaning. This technique improves information retrieval by reducing the size of indexing files.

Natural language processing (NLP)

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

- **Summarization:** This technique provides a synopsis of long pieces of text to create a concise, coherent summary of a document’s main points.
- **Part-of-Speech (PoS) tagging:** This technique assigns a tag to every token in a document based on its part of speech—that is, denoting nouns, verbs, adjectives, and so on. This step enables semantic analysis on unstructured text.
- **Text categorization**: This task, which is also known as text classification, is responsible for analyzing text documents and classifying them based on predefined topics or categories. This sub-task is particularly helpful when categorizing synonyms and abbreviations.
- **Sentiment analysis**: This task detects positive or negative sentiment from internal or external data sources, allowing you to track changes in customer attitudes over time. It is commonly used to provide information about perceptions of brands, products, and services. These insights can propel businesses to connect with customers and improve processes and user experiences.

Information extraction

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

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

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

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

- **Data mining**

- Data mining is the process of identifying patterns and extracting useful insights from big data sets. This practice evaluates both structured and unstructured data to identify new information, and it is commonly utilized to analyze consumer behaviors within marketing and sales. Text mining is essentially a sub-field of data mining as it focuses on bringing structure to unstructured data and analyzing it to generate novel insights. The techniques mentioned above are forms of data mining but fall under the scope of textual data analysis.

Applications of text mining with social media

Text mining helps companies leverage the omnipresence of social media platforms/content to improve a business’s products, services, processes and strategies. Some of the most interesting use cases for social media text mining include the following:

- Customer insights and sentiment analysis: Social media text mining enables businesses to gain deep insights into customer preferences, opinions and sentiments. Using programming languages like Python with high-tech platforms like NLTK and SpaCy, companies can analyze user-generated content (e.g., posts, comments and product reviews) to understand how customers perceive their products or services. This valuable information helps decision-makers refine marketing strategies, improve product offerings and deliver a more personalized customer experience.

- Improved customer support: When used alongside text analytics software, feedback systems (like chatbots), net-promoter scores (NPS), support tickets, customer surveys and social media profiles provide data that helps companies enhance the customer experience. Text mining and sentiment analysis also provide a framework to help companies address acute pain points quickly and improve overall customer satisfaction.

- Enhanced market research and competitive intelligence: Social media text mining provides businesses a cost-effective way to conduct market research and understand consumer behavior. By tracking keywords, hashtags and mentions related to their industry, companies can gain real-time insights into consumer preferences, opinions and purchasing patterns. Furthermore, businesses can monitor competitors' social media activity and use text mining to identify market gaps and devise strategies to gain a competitive advantage.

- Effective brand reputation management: Social media platforms are powerful channels where customers express opinions en masse. Text mining enables companies to proactively monitor and respond to brand mentions and customer feedback in real-time. By promptly addressing negative sentiments and customer concerns, businesses can mitigate potential reputation crises. Analyzing brand perception also gives organizations insight into their strengths, weaknesses and opportunities for improvement.

- Targeted marketing and personalized marketing: Social media text mining facilitates granular audience segmentation based on interests, behaviors and preferences. Analyzing social media data helps businesses identify key customer segments and tailor marketing campaigns accordingly, ensuring that marketing efforts are relevant, engaging and can effectively drive conversion rates. A targeted approach will optimize the user experience and enhance an organization's ROI.

- Influencer identification and marketing: Text mining helps organizations identify influencers and thought leaders within specific industries. By analyzing engagement, sentiment and follower count, companies can identify relevant influencers for collaborations and marketing campaigns, allowing businesses to amplify their brand message, reach new audiences, foster brand loyalty and build authentic connections.

- Crisis management and risk management: Text mining serves as an invaluable tool for identifying potential crises and managing risks. Monitoring social media can help companies detect early warning signs of impending crises, address customer complaints and prevent negative incidents from escalating. This proactive approach minimizes reputational damage, builds consumer trust and enhances overall crisis management strategies.

- Product development and innovation: Businesses always stand to benefit from better communication with customers. Text mining creates a direct line of communication with customers, helping companies gather valuable feedback and uncover opportunities for innovation. A customer-centric approach enables companies refine to existing products, develop new offerings and stay ahead of evolving customer needs and expectations.

Reference:

- https://www.ibm.com/topics/text-mining