Module 5

Text Mining, Naïve Bayes Analysis, Support Vector Machine, Web Mining and Social Network Analysis

Text Mining
• Text mining is the art and science of discovering knowledge, insights, and patterns from an organized collection of textual databases.
• Textual mining can help with frequency analysis of important terms and their semantic relationships.
• Text is an important part of the growing data in the world. Social media technologies have enabled users to become producers of text and images and other kinds of information.
• Text mining can be applied to large-scale social media data for gathering preferences and measuring emotional sentiments.
• It can also be applied to societal, organizational, and individual scales.

Text Mining Examples
Text mining works on texts from practically any kind of sources from any business domains, in any formats, including Word documents, PDF files, XML files, and so on. Here are some representative examples:
1. In the legal profession
2. In academic research
3. In the world of finance
4. In medicine
5. In marketing
6. In the world of technology and search

Text Mining Applications
Marketing: The voice of the customer can be captured in its native and raw format and then analyzed for customer preferences and complaints.
a. Social personas are a clustering technique to develop customer segments of interest. Consumer input from social media sources, such as reviews, blogs, and tweets, contain numerous leading indicators that can be used toward anticipating and predicting consumer behavior.
b. A “listening platform” is an application, which in real time, gathers social media, blogs, and other textual feedback, and filters out the chatter to extract true consumer sentiment. The insights can lead to more effective product marketing and better customer service.
c. The customer call center data can be analyzed for patterns of customer complaints. Decision trees can organize this data to create decision choices that could help with product management activities and to

become proactive in avoiding those complaints.

Text Mining Applications

Business operations:

a. Social network analysis and text mining can be applied to emails, blogs, social media and other data to measure the
emotional states and the mood of employee populations. Sentiment analysis can reveal early signs of employee dissatisfaction and this then can be proactively managed.

b. Studying people as emotional investors and using text analysis of the social Internet to measure mass psychology can help in obtaining superior investment returns.

Text Mining Applications

Legal: In legal applications, lawyers and paralegals can more easily search case histories and laws for relevant documents in a particular case to improve their chances of winning.

a. Text mining is also embedded in e-discovery platforms that helps in the process of sharing legally mandated documents.

b. Case histories, testimonies, and client meeting notes can reveal additional information, such as comorbidities in a health care situation that can help better predict high-cost injuries and prevent costs.

Text Mining Applications

Governance and politics: Governments can be overturned based on a tweet from a self-immolating fruit-vendor in Tunisia.

a. Social network analysis and text mining of large scale social media data can be used for measuring the
emotional states and the mood of constituent populations. Microtargeting constituents with specific messages gleaned from social media analysis can be a more efficient use of resources.

b. In geopolitical security, Internet chatter can be processed for realtime information and to connect the dots on any emerging threats.

Text Mining Process

Text mining is a semi-automated process. Text data needs to be gathered, structured, and then mined, in a three-step process

1. The text and documents are first gathered into a corpus and organized.
2. The corpus is then analyzed for structure. The result is a matrix mapping important terms to

source documents.
3. The structured data is then analyzed for word structures, sequences, and frequency.

Text Mining Process

Text Mining Process
Term-document matrix (TDM): This is the heart of the structuring process. Free flowing text can be transformed into numeric data, which can then be mined using regular data mining techniques.
1. The technique used for structuring the text is called the bag-ofwords technique. This approach measures the frequencies of select important words and/or phrases occurring in each document. This creates a **t × d**, term-by-document matrix (TDM), where t is the number of terms and d is the number of documents.
2. Creating a TDM requires making choices of which terms to include. The terms chosen should reflect the stated purpose of the text mining exercise. The bag of words should be as extensive as needed, but should not include unnecessary stuff that will serve to confuse the analysis or slow the computation.
Text Mining Process

Text Mining Process
Here are some considerations in creating a TDM.
1. A large collection of documents mapped to a large bag of words will likely lead to a very sparse matrix if they have few common words. Reducing dimensionality of data will help improve the speed of analysis and meaningfulness of the results. Synonyms, or terms with similar meaning, should be combined and should be counted together, as a common term. This would help reduce the number of distinct terms of words or "tokens."
2. Data should be cleaned for spelling errors. Common spelling errors should be ignored and the terms should be combined. Uppercase–lowercase terms should also be combined.
3. When many variants of the same term are used, just the stem of the word would be used to reduce the number of terms. For instance, terms like customer order, ordering, and order data should be combined into a single token word, called "order."

Text Mining Process
4. On the other side, homonyms (terms with the same spelling but different meanings) should be counted separately. This would enhance the quality of analysis. For example, the term order can mean a customer order, or the ranking of certain choices. These two should be treated separately. "The boss ordered that the customer data analysis be presented in chronological

order." This statement shows three different meanings for the word "order." Thus, there will be a need for a manual review of the TDM.
5. Terms with very few occurrences in very few documents should be eliminated from the matrix. This would help increase the density of the matrix and the quality of analysis.
6. The measures in each cell of the matrix could be one of several possibilities. It could be a simple count of the number of occurrences of each term in a document. It could also be the log of that number. It could be the fraction number computed by dividing the frequency count by the total number of words in the document. Or there may be binary values in the matrix to represent whether a term is mentioned or not. The choice of value in the cells will depend upon the purpose of the text analysis.

Comparing Text Mining and Data Mining

Dimension	**Text Mining**	**Data Mining**
Nature of data	Unstructured data: words, phrases, sentences	Numbers; alphabetical and logical values
Language used	Many languages and dialects used in the world; many languages are extinct, new documents are discovered	Similar numerical systems across the world
Clarity and precision	Sentences can be ambiguous; sentiment may contradict the words	Numbers are precise
Consistency	Different parts of the text can contradict each other	Different parts of data can be inconsistent, thus, requiring statistical significance analysis
Sentiment	Text may present a clear and consistent or mixed sentiment, across a continuum. Spoken words adds further sentiment	N/A
Quality	Spelling errors. Differing values of proper nouns, such as names. Varying quality of language translation	Issues with missing values, outliers, and so on
Nature of analysis	Keyword-based search; coexistence of themes; sentiment mining	A full wide range of statistical and machine-learning analysis

		for relationships and differences

Naïve Bayes Analysis
• Naïve Bayes technique is a is supervised machine learning technique that that uses probability theory based analysis.
• It is machine learning technique that computes the probabilities of an instance of belonging to each of many target classes, given the prior probabilities of classification using individual factors.

What is Support Vector Machine?
• "Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges.
• However, it is mostly used in classification problems.
• In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate.
• Then, we perform classification by finding the hyperplane that differentiate the two classes very well (look at the below snapshot).

What is Support Vector Machine?

How does it work?
Thumb rule to identify the right hyper-plane
• Select the hyper-plane which segregates the two classes better
• Maximizing the distances between nearest data point (either class) and hyper-plane. This distance is called as **Margin**.

SVM Model
• $f(x) = W.X + b$
• W is the normal to the line, X is input vector and

b the bias
• W is known as the weight vector

SVM Model

Advantages of SVM
• The main strength of SVM is that they work well even when the number of SVM features is much larger than the number of instances.
• It can work on datasets with huge feature space, such is the case in spam filtering, where a large number of words are the potential signifiers of a message being spam.
• Even when the optimal decision boundary is a nonlinear curve, the SVM transforms the variables to create new dimensions such that the representation of the classifier is a linear function of those transformed dimensions of the data.
• SVMs are conceptually easy to understand. They create an easy-to-understand linear classifier. By working on only a subset of relevant data,. they are computationally efficient. SVMs are now available with almost all data analytics toolsets.

Disadvantages of SVM
• The SVM technique has two major constraints
– It works well only with real numbers, i.e., all the data points in all the dimensions must be defined by numeric values only,
– It works only with binary classification problems. One can make a series of cascaded SVMs to get around this constraint.
• Training the SVMs is an inefficient and time consuming process, when the data is large.
• It does not work well when there is much noise in the data, and thus has to compute soft margins.
• The SVMs will also not provide a probability estimate of classification, i.e., the confidence level for classifying an instance.

Web Mining
• Web mining is the art and science of discovering patterns and insights from the World Wide Web so as to improve it.
• The World Wide Web is at the heart of the digital revolution. More data is posted on the Web every day than

was there on the whole Web just 20 years ago. Billions of users are using it every day for a variety of purposes.
• The Web is used for ecommerce, business communication, and many other applications.
• Web mining analyzes data from the Web and helps find insights that could optimize the web content and improve the user experience.
• Data for web mining is collected via web crawlers, web logs, and other means.

Web Mining

Here are some characteristics of optimized websites:
1. Appearance: Aesthetic design; well-formatted content, easy to scan and navigate; and good color contrasts.
2. Content: Well-planned information architecture with useful content; fresh content; search-engine optimized; and links to other good sites.
3. Functionality: Accessible to all authorized users; fast loading times; usable forms; and mobile enabled.

Web Mining

• The analysis of web usage provides feedback on the web content and also the consumer's browsing habits. This data can be of immense use for commercial advertising, and even for social engineering.
• The Web could be analyzed for its structure as well as content.
• The usage pattern of web pages could also be analyzed.
• Depending upon objectives, web mining can be divided into three different types: web usage mining, web content mining, and web structure mining

Web Mining Structure

Web Content Mining

• A website is designed in the form of pages with a distinct URL (universal resource locator). A large website may contain thousands of pages. Those pages and their content are managed using content management systems.
• Every page can have text, graphics, audio, video, forms, applications, and

more kinds of content, including user-generated content.
• The websites make a record of all requests received for its page/URLs.
• The log of these requests could be analyzed to gauge the popularity of those pages.
• The textual and application content could be analyzed for its usage by visits to the website.
• The pages on a website themselves could be analyzed for quality of content.
• The unwanted pages could be transformed with different content and style, or they may be deleted altogether. Similarly, more resources could be assigned to keep the more popular pages more fresh and inviting.

Web Structure Mining

The Web works through a system of hyperlinks using the hypertext transfer protocol (http).
The structure of web pages could also be analyzed to examine the structure of hyperlinks among pages.
There are two basic strategic models for successful websites: hubs and authorities.
1. Hubs: The pages with a large number of interesting links would serve as a hub, or a gathering point, where people access a variety of information. Media sites like Yahoo.com or government sites would serve that purpose. There are focused hubs like Traveladvisor.com and many websites which could aspire to become hubs for new emerging areas.
2. Authorities: Ultimately, people would gravitate toward pages that provide the most complete and authoritative information on a particular subject, including user reviews. These websites would have the most number of inbound links. Thus, Mayoclinic.com would serve as an authoritative page for expert medical opinion.

Web Usage Mining

• As a user clicks anywhere on a web page or application, the action is recorded by many entities in many locations. The browser at the client machine will record the click, and the web server providing the content would also log onto the pages-served activity. The entities between the client and the server, such as the router, proxy server, or ad server, too, would record that click.
• The goal of web usage is to extract useful information from data generated through web page visits and transactions.
The activity data comes from data stored in server access logs, referrer logs, agent logs, and clientside cookies. The user characteristics and usage profiles are also gathered

directly, or indirectly, through syndicated data. Further, metadata, such as page attributes, content attributes, and usage data, are also gathered.

Web Usage Mining

The web content could be analyzed at multiple levels.

1. The server side analysis would show the relative popularity of the web pages accessed. Those websites could be hubs and authorities.
2. The client-side analysis could focus on the usage pattern or the actual content consumed and created by users.

a) Usage pattern could be analyzed using "clickstream" analysis, that is, analyzing web activity for patterns of sequence of clicks, and the location and duration of visits on websites.

b) Textual information accessed on the pages retrieved by users could be analyzed using text mining techniques. The text would be gathered and structured using the bag-of-words technique to build a term-document matrix. This matrix could then be mined using cluster analysis and association rules for patterns, such as popular topics, user segmentation, and sentiment analysis

Web Usage Mining Architecture

Web Mining Algorithms

• Hyperlink-Induced Topic Search (HITS) is a link analysis algorithm that rates web pages as being hubs or authorities.
• The most famous and powerful of HITS based algorithm is the PageRank algorithm.
• Invented by Google co-founder Larry Page, this algorithm is used by Google to organize the results of its search function.
• This algorithm helps determine the relative importance of any particular web page by counting the number and quality of links to a page.
• The websites with more number of links, and/or more links from higher-quality websites, will be ranked higher.
• It works similar to determining the status of a person in a society of people. Those with relations to more people and/or relations to people of higher status will be accorded a higher status.