*11* Text Mining

Learning Objectives

* Understand the concept of Text Mining and its many applications
* Know the process of text mining, including the concept of Term-Document Matrix and how it can be mined
* Compare between text mining and data mining
* Learn some important best practices of text mining

### INTRODUCTION

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.

#### Caselet: WhatsApp and Private Security

*Do you think whether what you post on social media remains private or not? Think again. A new dashboard shows how much personal information is out there, and how companies can construct ways to make use of it for commercial benefits. Here is a dashboard of conversations between two people Jennifer and Nicole over 45 days.*

*There is a variety of categories that Nicole and Jennifer speak about such as computers, politics, laundry, desserts. The polarity of Jennifer’s thoughts and tone is overwhelmingly positive, and Jennifer responds to Nicole much more than vice versa, identifying Nicole as the influencer in their relationship.*

*The data visualization reveals the waking hours of Jennifer, showing that she is most active around 0800 hours and heads to bed around midnight. 53 percent of her conversation is about food – and 15 percent about desserts. Maybe she’s a strategic person to push restaurant or weight-loss ads.*

*The most intimate detail exposed during this conversation is that Nicole and Jennifer discuss right-wing populism, radical parties, and conservative politics. It*

*exemplifies that the amount of private information obtained from your WhatsApp conversations is limitless and potentially dangerous.*

*WhatsApp is the world’s largest messaging service that has over 450 million users. Facebook recently bought this three-year-old company for a whopping $19 billion. People share a lot of sensitive personal information on WhatsApp that they may not even share with their family members.*

*(Sources: What Facebook Knows About You From One WhatsApp Conversation, by Adi Azaria, on LinkedIn, April 10, 2014).*

1. *What are the business and social implications of this kind of analysis?*
2. *Are you worried? Should you be worried?*

Text mining works on texts from practically any kind of sources from any business or non-business domains, in any formats including Word documents, PDF files, XML files, text messages, etc. Here are some representative examples

* In the *legal profession*, text sources would include law, court deliberations, court orders, etc.
* In *academic research*, it would include texts of interviews, published research articles, etc.
* The world of *finance* will include statutory reports, internal reports, CFO statements, and more.
* In *medicine*, it would include medical journals, patient histories, discharge summaries, etc.
* In *marketing*, it would include advertisements, customer comments, etc.
* In the world of *technology and search*, it would include patent applications, the whole of information on the World WideWeb, and more.

### TEXT MINING APPLICATIONS

Text mining is a useful tool in the hands of chief knowledge officers to extract knowledge relevant to an organization. Text mining can be used across industry sectors and application areas, including decision support, sentiment analysis, fraud detection, survey analysis, and many more.

*Marketing* The voice of the customer can be captured in its native and raw format and then analyzed for customer preferences and complaints.

1. Social personas are a clustering technique to develop customer segments of interest. Consumer input from social media sources, such as reviews, blogs, and tweets, contains numerous leading indicators that can be used towards anticipating and predicting consumer behavior.
2. A ‘listening platform’ is a text mining application, that in real-time, gathers social media, blogs, and other textual feedback, and filters out the chatter to extract true consumer sentiments. The insights can lead to more effective product marketing and better customer service.
3. The customer call center conversations and records 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.

*Business Operation* Many aspects of business functioning can be accurately gauged from analyzing text.

1. 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 which can then be proactively managed.
2. Studying people as emotional investors and using text analysis of the social Internet to measure mass psychology can help in obtaining superior investment returns.

*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.

1. Text mining is also embedded in e-discovery platforms that help in minimizing risk in the process of sharing legally mandated documents.
2. Case histories, testimonies, and client meeting notes can reveal additional information, such as morbidities in healthcare situations that can help better predict high-cost injuries and prevent costs.

*Governance and Politics* Government can be overturned based on a tweet originating from a self-immolating fruit vendor in Tunisia.

1. 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 when fighting democratic elections.
2. In geopolitical security, internet chatter can be processed for real-time information and to connect the dots on any emerging threats.
3. In academics, research streams could be meta-analyzed for underlying re- search trends.

### TEXT MINING PROCESS

Text mining is a rapidly evolving area of research. As the amount of social media and other text data grows, there is a need for efficient abstraction and categorization of meaningful information from the text.

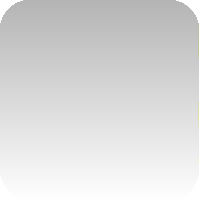
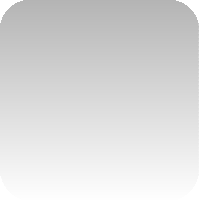
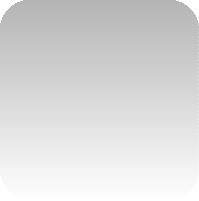
The first level of analysis is identifying frequent words. This creates a bag of important words. Texts – documents or smaller messages – can then be ranked on how they match to a particular bag-of-words. However, there are challenges with this approach. For example, the words may be spelled a little differently or there may be different words with similar meanings.

The next level is identifying meaningful phrases from words. Thus ‘ice’ and ‘cream’ will be two different keywords that often come together. However, there is a more meaningful phrase by combining the two words into ‘ice cream’. There might be similarly meaningful phrases like ‘Apple Pie’.

The next higher level is that of Topics. Multiple phrases can be combined into a Topic area. Thus, the two phrases above can be put into a common basket, and this bucket is called ‘Desserts’.

Text mining is a semi-automated process. Text data needs to be gathered, structured, and then mined, in a 3-step process (Figure 11.1)

Establish the Corpus of Text Gather documents,



clean and prepare for analysis

Structure using Term Document Matrix (TDM) Select a bag of words, compute frequencies of occurrence

Mine TDM for Patterns Apply data

mining tools like classification and cluster analysis

FIGURE 11.1 Text Mining Architecture

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.

### TERM DOCUMENT MATRIX

This is the heart of the structuring process. Free-flowing text can be transformed into numeric data in a TDM, which can then be mined using regular data mining techniques.

* There are several efficient techniques for identifying key terms from a text. There are less efficient techniques available for creating topics out of them. For this discussion, one could call keywords, phrases, or topics a term of interest. This approach measures the frequencies of

select important terms occurring in each document. This creates a

Term-by-Document Matrix (TDM), where is the number of terms and is

the number of documents (Table 11.1).

* 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 list of terms should be as extensive as needed but should not include unnecessary stuff that will serve to confuse the analysis or slow the computation.

Table 11.1 Term-Document Matrix

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Term Document Matrix | | | | | |
| Document/Terms | Investment | Profit | Happy | Success | … |
| Doc 1 | 10 | 4 | 3 | 4 |  |
| Doc 2 | 7 | 2 | 2 |  |  |
| Doc 3 |  |  | 2 | 6 |  |
| Doc 4 | 1 | 5 | 3 |  |  |
| Doc 5 |  | 6 |  | 2 |  |
| Doc 6 | 4 |  | 2 |  |  |
| … |  |  |  |  |  |

Here are some considerations in creating a TDM

* 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 the dimensionality of data will help improve the speed of analysis and meaningfulness of the results. Synonyms or terms with similar meanings 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’.
* 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.
* 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, order data, should be combined into a single token word, called ‘order’.
* On the other side, homonyms (terms with the same spelling but different meanings) should be counted separately. This would enhance the quality of the 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 orders data analysis be presented in chronological order”. This statement shows three different meanings of the word ‘order’. Thus, there will be a need for a manual review of the TD matrix.
* Terms with very few occurrences in the documents should be eliminated from the matrix. This would help increase the density of the matrix and the quality of the analysis.
* 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.

At the end of this analysis and cleansing, a well-formed, densely populated, rectangular, TDM will be ready for analysis. The TDM can be mined using all the available data mining techniques.

### MINING THE TDM

The TDM can be mined to extract patterns/knowledge. A variety of techniques could be applied to the TDM to extract new knowledge.

A simple application is to visualize the highest frequency terms. This can be done very attractively and colorfully in the form of a ‘wordcloud’. The wordcloud can be created after removing common words like prepositions. It can be done for the top *n* words such as top 100 words, to focus on the key terms used in the document. The attached wordcloud represents the speech by US President Barack Obama on the topic of terrorism.

Predictors of desirable terms could be discovered through predictive techniques, such as regression analysis. Suppose the word profit is a desirable word in a document. The number of occurrences of the word profit in a document could be regressed against many other terms in the TDM. The relative strengths of the coefficients of various predictor variables would show the relative impact of those terms on creating a profit discussion.



FIGURE 11.2 Wordcloud of Top 100 Words from US President’s Speech

Predicting the chances of a document being liked is another form of analysis. For example, important speeches made by the CEO or the CFO to investors could be evaluated for quality. If the classification of those documents (such as good or poor speeches) was available, then the terms of TDM could be used to predict the speech class. A decision tree could be constructed that makes a simple tree with a few decision points that predict the success of a speech 80 percent of the time. This tree could be trained with more data to become better over time.

Clustering techniques can help categorize documents by common profile. For example, documents containing the words investment and profit more often could be bundled together. Similarly, documents containing the words, customer orders, and marketing, more often could be bundled together. Thus, a few strongly demarcated bundles could capture the essence of the entire TDM. These bundles

could thus help with further processing, such as handing over select documents to others for legal discovery.

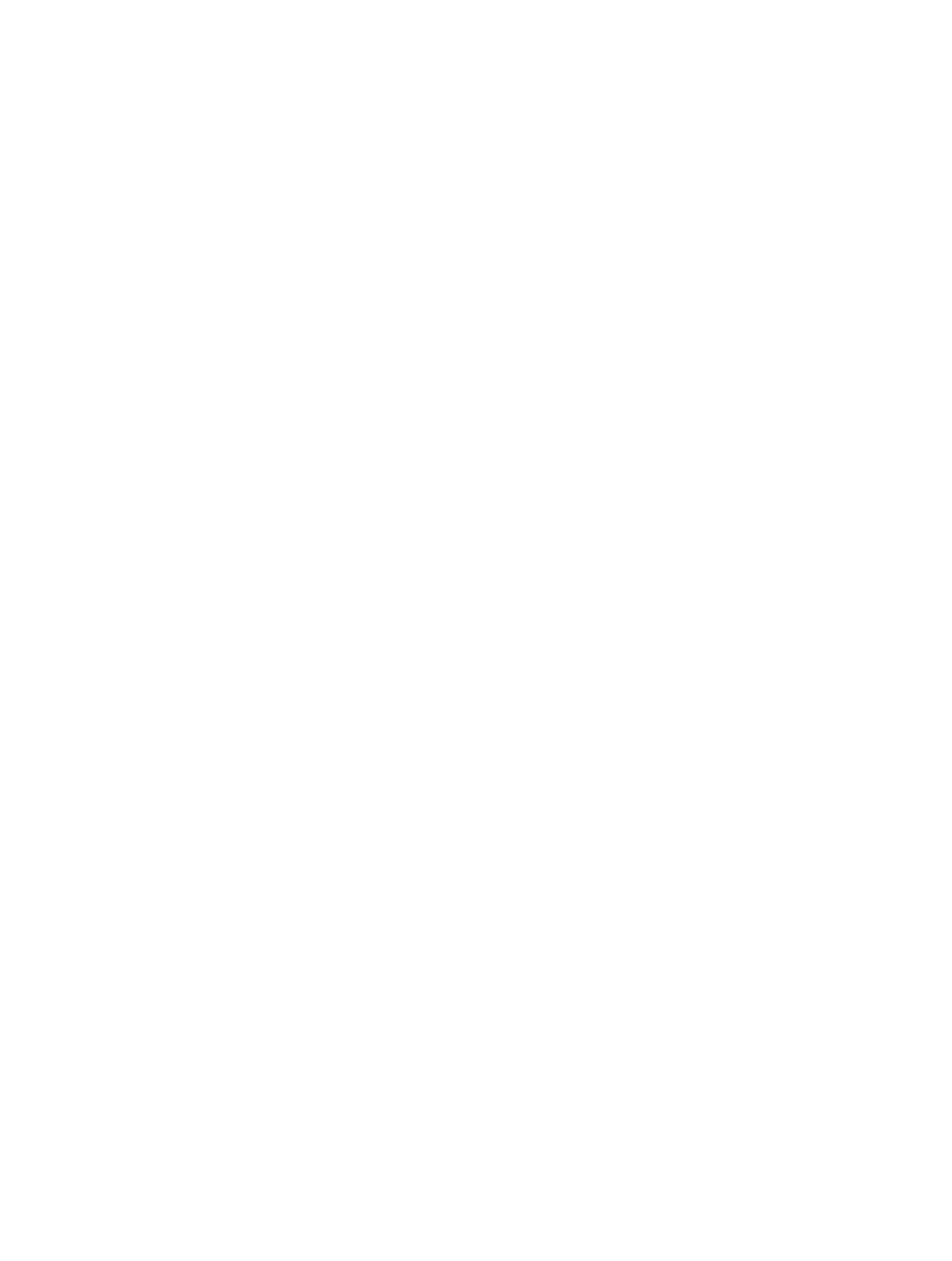
Association rule analysis could show relationships of coexistence. Thus, one could say that the words, tasty and sweet, occur together often (say 5 percent of the time); and further, when these two words are present, 70 percent of the time, the word happy, is also present in the document.

### COMPARING TEXT MINING AND DATA MINING

Text Mining is a form of data mining. There are many common elements between Text and Data Mining. However, there are some differences (Table 11.2). The key difference is that text mining requires the conversion of text data into frequency data before data mining techniques can be applied.

Table 11.2 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,    analysis |
| Sentiment | Text may present a clear and consistent or mixed sentiment, across a continuum. Spoken words add a further sentiment | Not applicable |
| Quality | Spelling errors. Differing values of proper nouns such as names. Varying quality of language translation | Issues with missing values, outliers, etc. |
| Nature of Analysis | Keyword-based search; co-existence of themes; sentiment mining; | A full wide range of statistical and machine learning analysis for relationships and differences |



### TEXT MINING BEST PRACTICES

Many of the best practices that apply to the use of data mining techniques will also apply to text mining.

* The first and most important practice is to ask the right question. A good question is the one that gives an answer and would lead to large payoffs for the organization. The purpose and the key question will define how and at what levels of granularity the TDM would be made. For example, TDM defined for simpler searches would be different from those used for complex semantic analysis or network analysis.
* A second important practice is to be creative and open in proposing imaginative hypotheses for the solution. Thinking outside the box is important, both in the quality of the proposed solution as well as in finding the high-quality datasets required to test the hypothesized solution. For example, a TDM of consumer sentiment data should be combined with customer order data to develop a comprehensive view of customer behavior. It’s important to assemble a team that has a healthy mix of technical and business skills.
* Another important element is to pursue the problem iteratively. Too much data can overwhelm the infrastructure and also befuddle the mind. It is better to divide and conquer the problem with a simpler TDM, with fewer terms and fewer documents and data sources. Expand as needed, in an iterative sequence of steps. In the future, add new terms to help improve predictive accuracy.
* A variety of data mining tools should be used to test the relationships in the TDM. Different decision tree algorithms could be run alongside cluster analysis and other techniques. Triangulating the findings with multiple techniques, and many what-if scenarios help build confidence in the solution. Test the solution in many ways before committing to deploy it.

## Conclusion

Text Mining is diving into the unstructured text to discover valuable insights about the business. The text is gathered and then structured into a term-document matrix based on the frequency of a bag of words in a corpus of documents. The TDM can then be mined for useful, novel patterns, and insights. While the technique is important, the business objective should be well understood and should always be kept in mind.

## Questions

1. Why is text mining useful in the age of social media?
2. What kinds of problems can be addressed using text mining?
3. What kinds of sentiments can be found in the text?

*Do a text mining analysis of sales speeches by three salesmen.*

1. Did you know your team can build PowerPoint muscles? Yes, I help build PowerPoint muscles. I teach people how to use PowerPoint more effectively in business. Now, for instance, I’m working with a global consulting firm to train all their senior consultants to give better sales presentations so they can close more business.
2. I train people on how to make sure their PowerPoint slides aren’t a complete disaster. Those who attend my workshop can create slides that are 50 percent more clear and 50 percent more convincing by the end of the training, based on scores students give each other before and after the workshop. I’m not sure if my training could work at your company. But I’d be happy to talk to you about it.
3. You know how most business people use PowerPoint but most use it pretty poorly? Well, bad PowerPoint has all kinds of consequences – sales that don’t close, good ideas that get ignored, time wasted building slides that could have been used developing or executing strategies. My company shows businesses how to use PowerPoint to capture those sales, bring attention to those great ideas, and use those wasted hours on more important projects.

The purpose is to select the best speech.

1. How would you select the right bag of words?
2. If speech #1 was the best speech, use the TDM to create a rule for good speeches.

## True/False

1. Text mining helps determine patterns within the written text.
2. A corpus is the collection of all the documents selected for text mining.
3. TDM refers to TermDocument Method.
4. All data mining techniques can be applied to the TDM.
5. Creating a TDM involves creating a matrix for all the terms and words in the document.
6. Text sentiment analysis can reveal early signs of employee dissatisfaction.