BARRY DEVILLE
GURPREET SINGH BAWA

TEXT AS DATA

COMPUTATIONAL METHODS OF UNDERSTANDING WRITTEN EXPRESSION USING SAS

WILEY

CHAPTER 1

Text Mining and Text Analytics

This chapter describes some of the background and recent history of text analytics and provides real-world examples of how text analytics works and solves business problems. This treatment provides examples of common forms of text analytics and examples of solution approaches. The discussion ranges from a history of the analytical treatment of text expression up to the most recent developments and applications.

BACKGROUND AND TERMINOLOGY

The analysis of written and spoken expression has been developing as a computer application over several decades. Some of the earliest research in machine learning and artificial intelligence dealt with the problem of reading and interpreting text as well as in text translation (machine translation). These early activities gave rise to a field of computer science known as *natural language processing (NLP)*. The recent rapid development of computer power – including processing power, large data, high bandwidth communication, and cloud-based, high-capacity computer memory – has provided a major new (and considerably broadened) emphasis on computerized text processing and text analysis.

TEXT ANALYTICS: WHAT IS IT?

Text processing and text analysis are components of the developing area of understanding written and spoken expression. Commonly occurring text documents – such as traditional newspapers, journals and periodicals, and, more recently, electronic documents, such as social media posts and emails – are forms of written expression. This active, multilayered area in current computer applications joins well-established, traditional fields such as linguistics and literary analysis to form the outline of the emerging field we call *text analytics*.

Current approaches to text analytics operate in two reinforcing directions that incorporate traditional forms of linguistic and literary analysis with a wide range of statistical, artificial intelligence (AI), and cognitive computing techniques to effectively process written and spoken expressions. The decoded expressions are used to drive

a wide range of computer-mediated inference tasks that includes artificial intelligence, cognitive computing, and statistical inference. An everyday example is when we speak or type in a destination in order to receive an optimal driving route. Similarly, a call center agent might decipher multiple forms of common requests in order to construct the most effective solution approach.

Our treatment throughout the chapters to come includes examples of common forms of text analytics and examples of solution approaches. The discussion ranges from a history of the analytical treatment of text expression up to the most recent developments and applications. Since speech is quickly becoming an important form of unstructured data, a final chapter takes up the topic of rendering speech to text.

Computer science and AI emerged as formal disciplines in the aftermath of World War II. An early application of computers to the analysis of written expression, natural language processing, took a universal approach, designed to apply regardless of what language the text was written in – English, Spanish, or Chinese. The techniques that have been developed also apply regardless of the source of the text to be analyzed. With the widespread availability of speech-to-text engines, it is also possible to consider a wide variety of spoken documents as potential sources for text analytics.

An important goal of NLP is to decompose text constructs (sentences, paragraphs, articles, chapters) into various kinds of entities, verbs, semantic constructs (like articles and conjunctions), and so on. The sentence "See Spot run" may be processed and encoded into an NLP representation as: declarative sentence (intransitive); Spot – Subject (Animal/Dog); run – Verb (motion).

Historically, NLP relied on various linguistic analysis capabilities, including extensive logical processing and reasoning capabilities. As computing capabilities have expanded, NLP has increasingly relied on a range of computational approaches to enhance the range of NLP results. An emerging area of NLP includes statistical natural language processing (SNLP). This form of NLP can be used to craft high-level representations of textual documents so that relationships between and among the documents can be computed statistically. The statistical capability also improves the accuracy of the NLP processing itself.

One recent area of written language processing includes statistical document analysis (SDA). Like SNLP, SDA enables us to show the statistical relationships between and among the various components of a textual document. Further, it enables us to summarize the document using multivariate statistical techniques like cluster analysis and latent class analysis. Predictive analytics such as regression analysis, decision trees, and neural networks can also be used.

As computer processing and storage have continued to grow, so too have a variety of deep learning applications. One such application is the Bidirectional Encoder Representations from Transformers (BERT), a deep-learning application for research at Google AI language.[i]

BERT can be leveraged for tasks such as categorization, entity extraction, and natural language generation. Deep learning approaches require significant computing power and training. As the area of text analytics continues to unfold, we will likely see how deep learning approaches complement the capabilities offered in traditional text analytics, which are less computationally intensive and more than adequate for a wide range of tasks.

The fields of *text mining* and *text analytics* are recent applied areas of SDA used in a variety of general-purpose social and economic settings. Text mining often refers to the construction of statistical or numerical models or predictions. Common sources of data include customer service logs and emails, customer use records for warranty issue analysis and defect detection. Text analytics often refers to semantically based applications – for example, customer analytics (who talks to whom and what do they say?), competitive analysis (brand metrics, mentions), and content management (the creation of taxonomies, web page characterization).

Brief History of Text

Language is a form of communication, and text is a written form of language. Text comes in a variety of symbolic forms. In addition to the alphabetic representation we see capturing the written expression in this text, there are other encoding systems such as syllabaries that capture spoken syllables and logograms that capture pictographic representations. Linguistics distinguishes between phonograms – which

Figure 1.1 Traffic sign in Cherokee syllabary, Tahlequah, Oklahoma. *Source:* Shot November 11, 2007. By Uyvsdi. License: Public Domain.

capture parts of words like syllables in written expression – and logograms – which capture entire concepts.

Figure 1.1 shows an example of a pictographic representation – the STOP sign itself – an alphabetic representation (in Latin script) that spells the word "STOP" and a syllabary – in this case, one used to record the Cherokee language.

One of the earliest true writing systems, dating to the third millennium BCE, was cuneiform, originally a pictographic writing system that eventually evolved into a variety of alphabetic representations. One intermediate form of simplified cuneiform was Old Persian. It included a semi-alphabetic syllabary, using far fewer wedge strokes than earlier Assyrian versions of cuneiform. It included a handful of logograms for frequently occurring words such as "god" and "king" (see Figure 1.2).

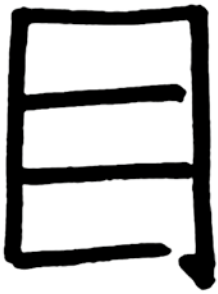
Chinese characters evolved in the second millennium BCE and, according to sources such as Dong,[ii] were first organized into a comprehensive writing system during the Qin dynasty (259–210 BCE).

Figure 1.2 Example of cuneiform recording the distribution of beer in southern Iraq, 3100–3000 BCE.
Source: BabelStone, Licensed under CC BY-SA 3.0.

These characters eventually gave rise to the widespread use of the characteristic logograms of Chinese in Asia (see Figure 1.3).

The representation of different writing systems is important for mapping language meanings between languages. Figure 1.4 shows a modern representation of the Chinese character for eye and the associated Latin script representation to show the translation between a pictograph (logogram) and syllabary.

Figure 1.3 Shang oracle bone script for character "Eye." Modern character is 目.
Source: Tomchen1989. Public Domain.



Mù

Figure 1.4 Modern Chinese representation of "eye" (mù).
Source: B. deVille.

Writing Systems of the World

Writing systems of the world that have evolved from ancient times to the present day can be organized into five categories[iii]: alphabets, abjads, abugidas, syllabaries, and logo-syllabaries.

1. **Alphabets.** Each letter represents a sound which can be either a consonant or a vowel. English uses an alphabet as do such related languages as French, German, and Spanish.
2. **Abjads.** Similar to alphabets except they are made up primarily of consonants. Vowel markings are absent or partial and may or may not be present. Hebrew and Arabic are the two main abjads in use today.
3. **Abugidas.** These are writing systems where consonant-vowel sequences are written as a unit. Consonants form the main units in the system and may stand alone or carry vowel notations with them. Abugidas evolved from a pre-Common Era Indian script called Brahmi and are prevalent in Southeast Asia.
4. **Syllabaries.** Here each character represents an entire syllable. A syllable is normally one consonant and one vowel. Japanese is an example.
5. **Logo-syllabary.** Each character can stand for a unique symbol or an entire word or idea. Chinese is an example.

Meaning and Ambiguity

Much of the work that we do in text mining – both hidden in the various text analytics engines we use as well as in the explicit user interventions we employ – will be directed at getting the best, most unambiguous meaning from the words or terms we use in the analysis. Numbers have relatively unambiguous properties and this facilitates their use in analytics. When we use *test*, however, it is normal to have a certain level of ambiguity in meaning. One of the main reasons for this is that textual terms are polysemous – one term may have multiple meanings. As an example of polysemy, think about the question, "Did you get it?" The question could be asking about understanding ("Yes, I understood!"), fetching an object ("I picked up the ladder this morning"), or receiving goods or services ("I got the vaccine last Tuesday").

Spoken and written forms of communication are prone to other breakdowns in communication. Figure 1.5 provides a rough illustration of the key features that are part of communication. One the earliest approaches to capturing and quantifying the information loss or gain contained in communication was the concept of entropy, formulated by Claude Shannon.[iv] Shannon borrowed the concept from thermodynamics and used it to rigorously engineer the communications properties of a range of communications methods and devices while working for Bell Labs. His contributions have placed him in the ranks of major figures in the establishment of the "computer age" along with such figures as Von Neuman, Alan Turing, Robert Noyce, Norbert Weiner, and Geoffrey Moore. As shown in the example in Figure 1.5, this approach is still used today.

The send–receive communications model reflects the notion of communication as capturing some kind of representation of an object that has been identified and passing the representation through various processing stages until the object representation has been received and decoded. In each of the processing stages, there is an opportunity for representation error to creep in so the message can degrade.

We can all informally observe the operation of entropy as we play the parlor game of passing a message from ear-to-ear in a circle of people. Words perfectly communicate when all the elements of the

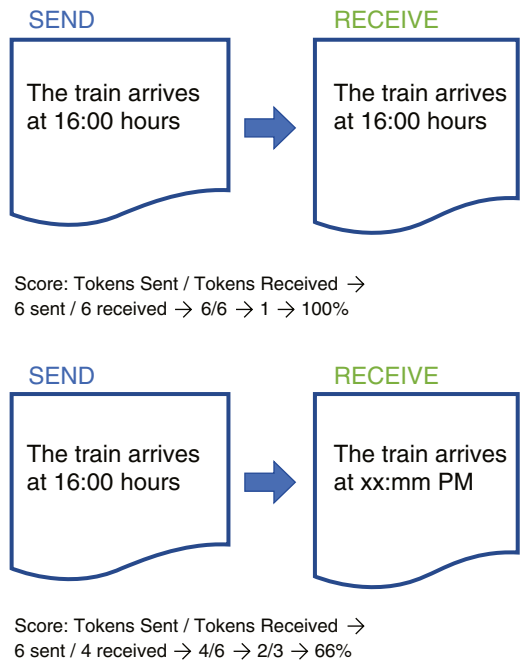


Figure 1.5 Encode–decode send–receive communications model.
Source: B. deVille.

sender's message are completely and accurately received and interpreted by the receiver.

Figure 1.5 provides an illustration of how entropy is calculated. When the full sentence (six tokens) is fully sent and received correctly there is 100 percent communication (zero entropy). When parts of the sentence are miscommunicated, for example, only four tokens are received; then communication drops to 66 percent. In this simple example, the first communication is better than the second communication (and there is an associated information gain of one-third, or over 33 percent).

Since text is a form of communication, we can gauge the accuracy and interpretation of the meaning of text using notions and measures of information entropy and information gain. Upon closer examination, we can also see that entropy and the statistical notion of correlation or association are related. The lower the entropy, the

higher the correlation. As we move more deeply into text analytics, the precision of textual meaning – sometimes reflected by low entropy and high association measures – becomes important, especially when we use text analytics to analyze large volumes of data. As with all data analysis tasks, the greater the accuracy of the analysis, the more useful the insights.

As a simple example of how we might calculate entropy, let's say we have documents about trains and boats. For purpose of illustration, we will use a radically simplified example in Table 1.1.

The "train" documents have the following words or tokens:

train wheels diesel track land

The "boat" documents have the following words or tokens:

boat rudder sail sea water

At this point, we can reframe the collection of text documents into a classification scheme that provides us with the ability to explore the communicative properties of words in a structured, reproducible fashion.

Table 1.1 Trains and Boats Example: Document Collection

		Component words/tokens							
Document	Class	wheels	rudder	diesel	sail	track	sea	land	water
1	TRAIN	x		x		x	x	x	x
2	BOAT		x		x				
3	TRAIN	x		x		x		x	x
4	BOAT	x			x		x	x	
5	TRAIN	x							
6	BOAT		x			x	x		x
7	TRAIN	x						x	
8	BOAT		x			x	x		
9	TRAIN	x							
10	BOAT			x	x		x		
11	TRAIN	x	x			x		x	x
12	BOAT	x		x			x	x	x

We can see that some terms/tokens appear in both kinds of documents: trains and boats. We can use entropy calculations to tell us which terms have the least entropy and are therefore most useful in classifying a document.

Shannon's formula for entropy (information theory) is . . .

$$\mathrm{H}(X) = -p\log_2(p) - q\log_2(q)$$

where H(X) is the expected value of the entropy calculation. It measures the difference in the probability of two outcomes, p and q.

In our example, p and q indicate whether the vehicle class is "train" or "boat." Logarithms have many useful properties, and the base 2 is used to support binary outcomes. This formula tells us that the expected entropy calculation of the features wheels through water in the above table will be formed by taking the logarithms of the probability of the features associated with one class – *train* – minus the probability of the features associated with the alternative class – *boat*.

If we calculate the entropy of the various terms in the example document (Table 1.2), we will see that the most useful term to unambiguously classify a document has the lowest entropy. This term is "sail." A high-entropy term like "diesel" is highly ambiguous, since it is applied to trains and boats with equal frequency.

Table 1.2 Entropy Calculation for Trains and Boats Example

Feature	Proportion (train)	Proportion (boat)	pr(train)	pr(boat)	Entropy
wheels	6/8	2/8	0.75	0.25	0.811
rudder	1/4	3/4	0.25	0.75	0.811
diesel	2/4	2/4	0.50	0.50	1
sail	0/3	3/3	0	1	0
track	3/5	2/5	0.60	0.40	0.971
sea	1/6	5/6	0.17	0.83	0.65
land	4/6	2/6	0.67	0.33	0.918
water	2/5	3/5	0.40	0.60	0.971

NOTES

i. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding Google AI Language* (Ithaca, NY: Cornell University: 2019). https://arxiv.org/abs/1810.04805v2.

ii. H. T.O. Dong, *A History of the Chinese Language* (London and New York: Routledge, 2014).

iii. F. Coulmas, *The Writing Systems of the World* (Hoboken, NJ: Wiley-Blackwell, 1919).

iv. J. J. Soni and R. Goodman, *A Mind at Play: How Claude Shannon Invented the Information Age* (New York: Simon & Schuster, 2017).