Start from the Question—A Guide to Unstructured Text Analysis

Seth Redmore is chief marketing officer at Lexalytics, Inc.
sredmore@lexalytics.com

Seth Redmore

Abstract

Text is one of the most important modes of human communication. To truly listen to your customers, partners, competitors, and employees, you must be able to understand the text they're exchanging with you and among themselves.

The science of unstructured text analytics has been around for many years and is used for many different purposes, from social marketing to customer experience management to helping managers understand customer support trends. It's used to help engineering ship new product features that match market needs.

This article discusses how to start from the question to understand what text you need to answer that question. It's a primer on several specific technologies that are used for text analytics, and explains different ways of implementing an unstructured text analytics system that will lead to the greatest probability of success.

Introduction

You've heard that 80 percent of the data in an enterprise is unstructured. This number actually comes from a "rule of thumb" that Merrill Lynch provided in the late 80s and hasn't been backed up by serious research[1]. I want you to completely ignore that number because it can lead to a pathological thought process that will keep you from getting real benefit from all the unstructured data you have collected or generated.

In other words, stop thinking about the data, at least at first, and start thinking about the question. This article

[1] https://en.wikipedia.org/wiki/Unstructured_data

will guide you through how to start from that question, then understand what unstructured (and structured) data you need to answer that question, and will then describe how you can work with your unstructured data to answer the question.

I will focus on unstructured text for this article, but with the exception of some specific technologies you'd bring to bear on a problem, the thinking works for any of the other types of unstructured data such as video or voice.

Start from the Question

I work for a text analytics vendor. We make our money turning unstructured text into structured data that can be fed into any of a large number of business analysis tools. I could want nothing more than to convince you that your text contains all the answers to any business question you might have, but that would be wildly dishonest.

Text is hard to work with and even native speakers of the same language have trouble understanding each other. A classic study by the University of Pittsburgh (Wiebe et al., 2005) followed grad students who were trained for 40 hours to classify sentence polarity for sentiment. They only agreed about 80 percent of the time.

Right off the bat, you're dealing with technologies that are inherently imperfect, and that will remain so for all time. As they say, "it is what it is," but this makes the following sentence all the more important:

What question do you want to answer?

Sometimes you don't know the right question and that's fine. You can start broadly with something like "Hey, we have a location that seems to be underperforming on sales. Can we see the comments from that location?" Note that you already used your structured data to determine that "something is up or down," namely your sales numbers.

Typical areas for questions involving unstructured text include:

- What is the discussion around my brand?
- What is the discussion around my competitor's brand?
- Why do I have a low net-promoter score? What can I do to improve it?
- What is the most common problem that customers are having with my product?
- Are my customers trying to reach me over social media? (Note that this isn't really an analytics question, it's more of a customer experience management question. You should probably know the answer to this one so you can start measuring it.)
- What do people like about our new advertising campaign?
- What are patients saying about their experience with our new drug?

The important thing is to ask not, "Whoa, I have all this text—what can I do with it?" but instead to ask, "What do I want to know?" It is perfectly fine to start with a broad question; just understand that the more specific you make it, the better your chance of actually getting an answer that is useful to your business.

What Data Do You Need?

Now you have a question. What data potentially contains the answer? Before we get into answering that, understand that over half of your time is going to be spent managing the content itself. Having clean, well-maintained data is more than half the battle of getting good results and will be an issue for you even with a fully off-the-shelf, outsourced SaaS system. There is no shortcut for this and you will need to continuously monitor the quality of the data coming in.

Having said that, there are four broad types of text content:

- Social media

- E-mail
- News
- Technical (patents, legal documents, research papers)

The choice of content is completely dictated by your question. You're not going to be looking in patents to see why more customers aren't coming into your San Francisco store.

Each of these content types presents challenges. You'll have to put up with them to get to your answer and it's good to go in with your eyes open.

Social Content

Social content can include the obvious, such as Twitter or Facebook, and it also includes message boards, YouTube comments, reviews, wikis, and blogs. There's a certain amount of crossover to "news" content with blogs—some are as influential as traditional news sources and some are just people shouting into the void.

Social content is generally the easiest content to get your hands on. The great volume produced daily—every single day—creates a problem. It's messy content. It's rife with misspellings, acronyms, in-jokes, emoji, spam, and other content not relevant to answering your question. The biggest issue with social content is filtering.

Other important challenges with social content include:

- Anonymization: It can be very hard to associate a tweet with an actual identity, one that carries demographic information or even shows that this is a real person with real opinions that you should count.
- Self-selection: Only people who want to share their opinion are going to share it.
- Demographic skew: According to Pew Research, (Duggan et al., 2015) Twitter users are 21 percent white, 27 percent black, and 25 percent Hispanic. Does this match your customer base?
- Bias toward consumer information: This might seem blindingly obvious, but B2B companies have tried in vain to mine Twitter for discussions about such things as dense wavelength division multiplexing. Yes, you'll get a few conversations about highly technical topics, but not many.

One interesting space is "enterprise social content." If your company has a wiki or is using a tool such as Slack or Yammer, it can be interesting to monitor that source for hotspots, but only if the question you're trying to answer has to do with "how are my projects going?"

E-mail

There are three main sources of e-mail messages for most non-government enterprises:

- Company internal
- External e-mail sent to your company—for example, to your support team
- E-mail sent from your company, particularly regarding regulations

You can generally associate an e-mail message with an identity. If it's coming into your support team, for example, you know who the customer is. If it's internal, you know the employee.

The major difficulty with e-mail comes in threading, duplication, and signatures. Do you want to triple count something that is part of a forwarded chain, or do you need technology to de-thread messages and only count them once? Signatures carry a variety of brand names and legalese that is uninteresting for your purposes and thus needs to be stripped out or ignored.

Another challenge with e-mail is the "expectation to be watched." If a customer sends you a complaint via e-mail, they're expecting that e-mail to be read and have implicitly given your permission to roll it up into a higher-level report. If two co-workers are complaining to each other about their boss, they might get upset if that boss comes to talk to them about that conversation—

even if it is completely within the company's legal right to watch that conversation.

This article is about analytics, so I won't spend time on data leakage prevention—which is a hot topic for regulated industries and applies most directly to communications (either social or e-mail) that are leaving a company.

News

News content is generally well written and can be clean. There's a bit of noise when you start pulling in blogs, and depending on your data service provider, you can get ancillary advertisements mixed in with the main content, which can throw off your analysis.

The biggest challenge with news content generally comes with the licensing terms. Are you allowed to keep the articles after analysis? For how long? Are you allowed to provide the full text to your analysts? These are questions you need to ask when working with news.

News articles tend to be information-dense and often contain information about topics you're not interested in. Think about how to zero in on what you specifically want to know.

Technical Content

This is a big bucket, and generally speaking the systems that handle technical content are specialized. Patents are often written to obfuscate and be as broad as possible. Research papers have specific terminology that may require special dictionaries to handle. Legal documents, such as contracts, have a strong combination of structured and unstructured information.

Access to technical content will vary with the type of content. Patents are generally easy to obtain but difficult to parse. You probably have all your contracts, so obtaining and processing copies of them is relatively simple. Research papers depend heavily on the publisher, and have many of the same licensing issues as news content has.

A Note on Personally Identifiable Information (PII)

I cannot offer legal advice, but if you are operating in the European Union, or analyzing content coming from the EU, make sure you understand your obligations with respect to privacy and deleting any social or other content that may contain personally identifiable information. Shifting laws are creating considerable confusion about the US/EU "Safe Harbor" program, but I expect it to settle down in the next year or so.

Theme/topic extraction helps you determine what all the buzz is about.

What Text Analytics Can Do

Unstructured text analysis (aka text mining or text analytics) uses several core technologies and terms that you should understand. These terms refer to the elements that are going to help you parse through all that data to answer your question. A few of the most important technologies follow:

Named entity extraction involves extracting the proper nouns being discussed, such as the names of companies, people, or products. Also known as named entity recognition, this feature helps you understand "who," "what," and "where." Typically named entities are proper nouns, such as the names of companies, people, or products. You extract them so you can associate other things with them (such as sentiment) or to get an early alert about new players in your space. If done well, named entity extraction can be a great discovery method for keeping on top of people moving around or staying informed about new companies and products entering your market.

Theme/topic extraction helps you determine what all the buzz is about. What are the common topics of conversation? Theme/topic extraction will tell you. This is automatic technology to extract key phrases and make key associations.

Categorization helps you sort content into "buckets." It's just the opposite of theme-topic extraction because

you have to configure your categories ahead of time—categories can include "buckets" such as "service" or "food" or "support."

Summarization isn't useful for things like tweets, but is valuable for analyzing 200-page research reports.

Sentiment analysis detects the tonality of the conversation. Some conversations have a polarity (positive/negative/neutral), others extend this to include analysis of other emotional axes, such as "frustrated/satisfied." Sentiment analysis can be performed at many levels; most common is "document sentiment." This analysis is only useful for short content. Generally you want to associate sentiment with entities or categories in order to zoom in on whatever was actually being discussed.

Intention extraction helps you determine if the person speaking "intends" to take a future action, such as buying or returning a product. It is a relatively new, unproven technology with little presence in the market. It shows promise for reducing customer churn and generating new business.

A Special Note on Machine Learning

There's recently been tremendous buzz around machine learning, particularly a subset called "deep learning." Machine learning is a foundational technology that does nothing by itself—you have to train it for a particular task. Most text mining uses machine learning.

Machine learning techniques are many: deep learning, conditional random field, maximum entropy, matrix factorization, and so on, but they all rely on having content to learn from that is similar to what you will be using to answer your question.

Sometimes machine learning and text analytics are presented as opposing techniques. In reality, they overlap. Almost all text mining tools use machine learning at some level and machine learning does things (such as image recognition) that aren't related to text mining at all. Which to use is simply a question of picking the best tool to answer your question.

If there's one thing you want to do (for example, "classify this content into buckets" such as determining the document sentiment or picking out the entities), then machine learning is an excellent tool. You'll need to have the content, and if you're performing supervised machine learning, you'll need to tag all the content. This can be a monumental task. There are also unsupervised machine-learning algorithms that look for patterns in the data and use them to make decisions about new data, but those are less common than the supervised machine-learning algorithms.

As an example, the output from a text analytics system (which itself is using machine learning) can be used as input to another predictive model built using your additional structured data. It is generally best to try to isolate the steps you're trying to take, meaning it will be harder to isolate issues and get meaningful results if you pour all your unstructured text right in with your structured data and try to get a prediction. Layering the problem—where the unstructured data processing leads to structured data that can be used with other structured data inside another machine learning model—will take you down a much more useful path.

Sometimes machine learning and text analytics are presented as opposing techniques. In reality, they overlap.

Understanding the Limitations of Text Analytics

Potential customers frequently ask text analytics vendors, "How accurate is your system?" We must explain that we can't answer the question because we haven't talked about what they're trying to do. The simple truth is that accuracy for any system will be heavily dependent on several factors:

- Specificity of the space: The broader you are looking, the less accurate you'll be, and the narrower your view, the better your results because context changes

between market spaces and words mean different things in different context.

- Cleanliness of the content: Dirty content means bad results.

- Inter-rater agreement: Most content is scored by multiple raters. If you have three raters for any given piece of content, and on average two of them agree, you have 66 percent inter-rater agreement. If your human raters don't agree whether something is positive or negative, how can you expect your machine to do well? Some content is inherently hard to rate, but generally speaking you need to make sure your coding rules are explicit, well thought out, and accompanied by plenty of examples for your raters.

- Amount of content: This isn't precisely about accuracy, but the more you have to work with, the better your results will be. Small datasets require the fancier technology—largeness has a benefit all its own.

Roughly speaking, most sentiment systems will be 65 to 75 percent accurate. Domain-specific tuning can push that accuracy higher, but beware of claims of 90 percent or better accuracy. Some languages are better than others—English has many linguistic resources, whereas Japanese is notoriously and provably hard to analyze accurately because of its related social norms.

It is important to think beyond accuracy. Instead, decide whether you are more concerned about precision or recall. If you aren't familiar with this concept, the difference is simple. Precision is "for the set of things you said were 'blue', how many actually were 'blue'?" Thus, if you have three blue things and one red thing, you have 75 percent precision. Recall is focused on "for the entire set of blue things in this bucket, how many blue things did you find?" If you find three blue things, and there were six blue things in the bucket, you have a recall of 50 percent.

Some applications are more sensitive to precision than recall—in other words, it's fine if you miss some content as long as you're sure that the content you received is absolutely correct. This is a common mindset for social listening applications. Others are more sensitive to recall—if you have two upset customers and you only get back to one, that's probably worse than thinking you have three upset customers, getting back to the two that were upset and not worrying about the third. In other words, making sure you don't miss anything.

More important than overall accuracy is "is it good enough for me to make a meaningful decision?" An extra few points of accuracy may cost you a month in configuration time but not lead to better business outcomes. Relating your decisions back to business results is more important than overall system "accuracy."

Buy vs. Integrate vs. Build

Once you have a good idea what your question is and what kind of data you need, you must determine how to gather and process the data to answer your question. There are four ways to approach this problem: fully integrated SaaS platforms, desktop tools, data warehouse integration, and fully customized systems. I'll discuss these in order of complexity, cost, and time.

It is important to think beyond accuracy. Instead, decide whether you are more concerned about precision or recall.

Fully Integrated SaaS Systems

To answer customer experience management questions or perform social marketing, you have many online vendors to choose from. They will acquire the content for you, process it using either their own technology or text analytics technology they've OEM'd, and present you with a set of visualizations, analysis, or alerts of your choice.

I won't mention any vendor names here, but I can offer suggestions about what to ask. Make sure the vendor you're considering can answer the questions you want

to ask. Make sure they have the data sources you need. Ask questions about reprocessing data—what happens when you change your configuration and you want to look back in time? Can you do that or are you limited to looking forward? Do you want to know the sentiment associated with individual entities or are you satisfied with document sentiment because you're dealing with short content? How much configuration or tuning are you interested in doing?

That last question is a bit tricky. Nobody wants to perform extensive configuration or tuning, but sometimes you simply must in order to ensure that the system recognizes your particular problem space. This could be a matter of gathering content and training, or it could be a matter of switching the meaning of a few sentiment phrases or configuring some search phrases as categories. As a rule of thumb, the more specific your space, the better your results for any given amount of work.

The major advantage of these SaaS systems is that they are complete solutions—data, storage, processing, and visualization. Some allow you to import your own private data, others don't.

Desktop Tools

If you have a relatively constrained set of content—such as a survey that you run periodically with fewer than one million responses—a desktop tool might be the best way to answer your question. These tools can range from add-ins for Microsoft Excel to full statistical packages. Most major stats packages contain some text analytics functionality. These work best if you have a limited number of analysts producing results to be distributed. Desktop works less well if you have a larger number of people who all want to view and manipulate the dashboards.

One indisputable advantage of desktop tools is that you can keep complete control of the data, so it can be the most secure option if you're dealing with highly confidential text.

These tools generally include all processing and visualization in one place, and you'll point them at a local dataset. You will need to "bring your own content," which can be quite a hassle, but if you're already collecting it, it's a much more tractable solution.

Data Warehouse Integration

Many larger organizations already have flourishing business intelligence data warehouses with associated storage and visualization tools. Often, they're already collecting text, and even if they aren't, the infrastructure exists for the storage and transformation of text.

If you are part of such an organization, it can be easiest to make use of the work that's already been done. You'll need to add the text analytics piece, and if it's not already there, the content that will answer your question. Most BI tools will work fine for visualizing the results of your text processing, and there are a myriad of free and open source tools in case you need a feature that isn't available in your chosen BI tool.

Fully Customized Systems

These are relatively rare beasts, but a handful of companies have decided they truly want to differentiate themselves on the basis of how they handle unstructured text, and so they spend $500,000 or more to build a complete system, including custom visualizations, data storage, content, etc.

If you are considering it, try something small first. Use a desktop tool. Use some of the online services. Talk to your data warehouse people—they can be really great. Only go this route if you are dead certain you know exactly what you're getting yourself into, as it is going to take longer and be more painful than you expect.[2]

You can get some truly different and spectacular results out of such a system, but you have to have a significant team and budget to be successful, and it will not be a one-time capital expense. It will require constant care and feeding.

Plan for Success

You have your question, you know the data that should answer your question, you know how you're going to

[2] https://en.wikipedia.org/wiki/Hofstadter%27s_law

actually implement the solution, and now you just need to get on with things.

- Start small. Answer the one question. You will learn a tremendous amount and your question might well change.
- Put time into testing, tuning, and verification. Make sure you can trust the results before you grow the system.
- Iterate frequently.
- Show value from answering the first question; only then should you grow the project.
- Constantly check your analytics program against your business goals.

These may seem like obvious, simple steps, but it is amazing how often we see companies bite off more than they can chew by starting too big. It is much easier to start small and grow up than to try to implement a large, unstructured project and scale it back. Starting small is less expensive, carries a lower risk, and can ramp easily to higher value. ■

References

Duggan, Maeve, Nicole B. Ellison, Cliff Lampe, Amanda Lenhart, and Mary Madden [2015]. "Demographics of Key Social Networking Platforms." http://www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms-2/

Wiebe, Janyce, Theresa Wilson, and Clare Cardie [2005]. "Annotating expressions of opinions and emotions in language," *Language Resources and Evaluation*, Vol. 39, No. 2–3, pp. 165-210.