



---

## **Text Analytics for Social Media**

Evolving Tools for an Evolving Environment

---



## Table of Contents

<b>Introduction .....</b>	<b>1</b>
<b>I. The Fundamentals of Text Analytics .....</b>	<b>2</b>
What Type of Document Is This and What Can We Do with It? .....	2
What Is This Document About? .....	2
What Are the Contents of This Document? .....	3
What Is the Author’s Attitude? .....	3
How Can We Manage and Use This Knowledge? .....	3
<b>II. Trusting Machines to Understand Humans .....</b>	<b>4</b>
<b>III. Needing Machines to Understand Humans .....</b>	<b>4</b>
Consistency .....	4
Objectivity .....	5
Subjectivity .....	5
Depth .....	5
Discovery .....	5
<b>IV. Social Media Makes Text Analytics a Necessity.....</b>	<b>6</b>
Reach Revisited .....	6
Trouble Spotting.....	6
Brand Affection Monitoring.....	7
Identifying Opinion Leaders .....	8
Monitoring One’s Impact on the Conversation .....	8
Trend Analysis and Predictive Modeling .....	9
<b>V. Text Analytics in the Future.....</b>	<b>9</b>
Twitter Reveals Moviegoers Partiality.....	9
Social Media Frightens Fast Food .....	10
Social Media Taps Hearts and Minds.....	12
<b>VI. Conclusion .....</b>	<b>13</b>

---

## About the author

Jim Sterne produced the world's first "Marketing on the Internet" seminar series in 1994. Today, Sterne is an internationally known speaker on electronic marketing and customer interaction and a consultant to FORTUNE Global 500® companies and Internet entrepreneurs. Sterne focuses his 25 years in sales and marketing on measuring the value of the Internet as a medium for creating and strengthening customer relationships. Sterne has written seven books, including:

*World Wide Web Marketing*, 3rd Edition

*Customer Service on the Internet*, 2nd Edition

*What Makes People Click: Advertising on the Web*

*E-mail Marketing*

*Web Metrics: Proven Methods for Measuring Web Site Success*

*Advanced Email Marketing*

*Social Media Metrics: How to Measure and Optimize Your Marketing Investment*

With a special focus on Web metrics, his company, Target Marketing ([targeting.com](http://targeting.com)), is dedicated to helping companies understand the possibilities and manage the realities of conducting business online. Sterne is the producer of the international eMetrics Marketing Optimization Summit ([emetrics.org](http://emetrics.org)) and is co-founder and current Chairman of the Web Analytics Association ([WebAnalyticsAssociation.org](http://WebAnalyticsAssociation.org)). Sterne was named one of the 50 most influential people in digital marketing by *Revolution*, the United Kingdom's premier interactive marketing magazine, and one of the top 25 Hot Speakers by the National Speakers Association.

---

## Introduction

---

Monitoring public conversations about a brand or a product requires heavy lifting from text analytics tools, and understanding the technical aspects of text analytics is a necessity for modern marketing executives.

This paper is for marketing professionals who want to understand the practical side of text analytics as a competitive advantage. This is neither a technical treatise nor a how-to handbook. It is a guide to a rapidly changing technology that can have a direct impact on your sales top line and financial bottom line.

Marketers love hard numbers - reach and frequency, website visits, multiple-choice questionnaire results - because anything quantifiable helps determine which marketing methods are working. We can crunch numbers into pleasing charts and graphs, and easily manipulate them to tell a truth that matches our preconceived notions.

But when faced with unstructured data - verbatim survey responses, written call center reports and e-mail messages - marketers cringe. Counting things is easy. Gleaning meaning and deciphering substance is daunting. The sheer volume of words makes the task appear insurmountable.

Text analytics strive to derive meaning from the written word. This is massively complex because human communication is so context dependent. Understanding a simple phrase like, "He saw the boy with the telescope," is tricky. Did he see the boy *through* the telescope or *carrying* the telescope?

In the movie *Broadcast News*, Holly Hunter tells Albert Brooks on the phone that they have to meet right away. Brooks replies, "Ok, I'll meet you at the place near the thing where we went that time." To an outsider this is meaningless, but anyone in a long-term relationship has had this experience.

Junior staff members can read through and code thousands of pages of customer comments to determine whether they are satisfied, if products are performing or if prospective customers are ripe for conversion. Unfortunately, they are the least experienced individuals in the organization and senior staff members simply do not have time to devote to this type of analysis.

Our inability to keep up with the volume of communication is compounded by the rise of social media. Our need to understand conversations taking place out in the sociosphere is critical. Blog posts, tweets, comments on YouTube and more reveal what is on the minds and in the hearts of the public. Absorbing these opinions is not humanly possible, so we turn to technology. Text analytics have been useful for decades and are growing more sophisticated.

Starting with a brief description of the underlying technologies from a marketer's perspective, we'll review what they do, how are they now being used in marketing and how they might be used tomorrow.

## I. The Fundamentals of Text Analytics

As tools become more capable and more available, we must learn to maximize their benefits. For a marketer to make the most of these technologies and work most effectively with technologists, it is necessary to be familiar with some basic concepts.

Text analytics attempt to build a subject-matter framework upon which we might hang more nuanced contexts. It starts by answering a few critical questions about the text we are trying to analyze.

### What Type of Document Is This and What Can We Do with It?

Document classification is the art of figuring out the nature of a document. Why? Because we want to treat a peer-reviewed, scientific journal article differently from a blog post or a tweet. An article includes references and citations and is cited by others. This helps us understand its relevance and provides an indication of relative importance.

A comment on a blog post is not an isolated passage and must be viewed in context with the original post, the comments that preceded it, the public discourse on the topic and the press coverage at the time of the post. While this is a seemingly simple task for automation, you need only look at the spotty effectiveness of a spam filter to realize the complexity and value of text analytics tools.

Identifying the type of document under consideration is a good first step and a requirement in moving on to the next step - subject matter.

### What Is This Document About?

Determining what a document is about is a challenge for Google. After identifying a document's subject matter, Google must determine how relevant it is to a given query. The results are often interesting, sometimes surprising and just useful enough to make us all Google addicts.

A number of analytical functions come into play, and you will hear any or all of these mentioned by analysts:

- Topic identification.
- Concept mining.
- Information extraction.
- Grammatical analysis.
- Natural language processing.
- Computational linguistics.

---

■ Analyzing unstructured, textual information is complex and overwhelming.

---

There are subtle differences in the above techniques that are significant to the technicians who work with them, but they are all aimed at trying to determine the subject of a document. All documents are about something in particular, but they can also be related to many other subjects. These tools identify the main subject while noting important secondary concepts.

Now that we know the topic, we need to know whether the document contains fact or fiction.

## What Are the Contents of This Document?

Known as semantic orientation, the goal of this step is to sort out the contents into buckets of fact, conjecture, opinion, belief stated as fact, etc.

Are the statements subjective or objective? “It seemed like hours until he came out,” indicates the observer’s state of mind more than on the actual passage of time. Lawyers spend years learning how to pick apart testimony in just this way when cross-examining witnesses.

With document type, subject matter and contents worked out, we are left wondering about the writer’s point of view.

## What Is the Author’s Attitude?

Sentiment analysis tries to identify the polarity of the content (positive, negative or neutral). “It seemed like hours,” could communicate unhappiness at having to wait or express delight that something wonderful went on and on. For example, “I love being put on hold for half an hour,” is clearly sarcasm to the human reader but must be explained to a sentiment analysis system.

## How Can We Manage and Use This Knowledge?

If a machine is going to learn, then context is needed so it can put the pieces together. The machine needs an ontology. An ontology is the sum of relationships and definitions within a subject that allows for machine manipulation. With context and meaning in hand, ontology management provides formal labeling and categorization of the information.

The computer systems we use must be taught the relationships between words. An ontology for fashion must include the association of sleeves to shirts, legs to pants and scarves, belts and jewelry to accessories.

Ontology management provides the handles that allow a computer to manipulate content in context. This is where we move from reporting about the content of a single document to analyzing trends across multiple documents and document types over time.

---

■ The volume of unstructured, textual information is growing shockingly fast.

---

As with most metrics, the trends are far more important than discrete scores when it comes to determining appropriate business actions. How much and how fast things change provide more value than static counts.

But can a machine really understand human communication? Isn't this what artificial intelligence experts have been seeking for decades?

---

## II. Trusting Machines to Understand Humans

---

Sentiment analysis is the most difficult task of all. Humans often have trouble understanding each other, even when speaking face to face. Without facial expressions or vocal clues, textual misunderstanding is common. That is why legal writing, for example, is excessively dense. It is an attempt to avoid misinterpretation. The problem of misinterpretation has challenged technology for years. In the past, enthusiastic experts were convinced of the power and sophistication that high-level math brought to bear on the problem.

The answer actually lies in humans working closely with machines. Computers can calculate that a given phrase is positive or negative with a certain degree of confidence. Humans with specific domain expertise can review low-confidence results and advise the machine how to grade them. Over time, the computer absorbs more and more of the expert's perspective and becomes more accurate and useful.

But even before the domain expert offers the first course correction to a text analytics system, the system has some distinct advantages that make it a necessity.

---

## III. Needing Machines to Understand Humans

---

Even if you were able to provide a nearly infinite number of intelligent, trained and dedicated humans, text analytics done by hand still has serious drawbacks.

### Consistency

---

Even the most diligent person will categorize things differently on any given day. Text analytics depend on always following the rules. Computers excel at rule following - they do as they're told - repeatedly and reliably. This is important because a good text analytics program incorporates a large, growing set of rules. If a human is able to follow all of the rules, all of the time, we call them "obsessive-compulsive."

---

■ Text analytics are now a necessary, competitive marketing tool.

---

## Objectivity

People suffer from biases rooted in culture, education, age, gender and other environmental factors. People are elastic; they learn, adapt and change. A computer will be faithful to the rules and yet, those rules on occasion can be tested, challenged and changed by humans to refine the results.

## Subjectivity

The most unbiased person on the planet is still human and therefore is capable of making mistakes. Seeing the word “hat” and interpreting it as “cat” is not easy to correct due to the inconsistency of the error and our inability to anticipate all of the factors that can lead to this error. People simply make mistakes.

## Depth

We sometimes have trouble finding the car keys. Even the smartest among us cannot remember everything - rules, definitions, cross-references and relationships. Text analytics for a given domain may embrace thousands of designations, hundreds of classifications and an untold number of situational definitions.

## Discovery

Inventors and artists have the unique ability to look at something and see it in a new or unique way. Einstein sees a clock tower and imagines it on a train, which triggers a line of reasoning that results in  $E=MC^2$ . Darwin sees different birds on different islands and formulates a means of evolution to explain it. Daniel Kirkwood identified a pattern in the relationship between the distance of a planet from the sun and its rotation time, which became known as Kirkwood's Law. Each example stands out as a unique event in scientific progress. We call them discoveries and they are rare.

Discoveries in text analytics are even more difficult due to the human problems of consistency, objectivity, subjectivity and depth, but computers have been designed and refined to discover patterns in data.

The torrent of text created daily in social media requires the process of monitoring and evaluating all that text to be automated. A text analytics engine can much more quickly uncover the relationships between terms of endearment for, or expressions of enmity against, a given brand or product and gauge the significance of any shift in those relationships.

---

■ Sentiment analysis is beyond the scope of humans or machines acting alone.

---



## IV. Social Media Makes Text Analytics a Necessity

Many-to-many communications have given the public enormous influential power. Public opinion is more trusted and respected than the promotional declarations of a company.

Text analytics are our only hope for monitoring, comprehending and participating in public discourse. While the variety of technologies may be perplexing, the uses for them are vital for calibrating the success of social media investments, and maintaining good relations and a competitive edge. From the advertiser's perspective, success is measured in the reach required to generate awareness.

### Reach Revisited

Counting how many people have been exposed to your brand has become much more complex than merely multiplying the number of commercials by the number of viewers. We can't just count the number of people who read something about McDonald's, Big Mac or other branded terms.

There are many other terms related to McDonald's (Mickey D's, Golden Arches, Ronald's, etc.) as well as movies referencing McDonalds (*Fast Food Nation* and *Super Size Me*), legal battles (the McLibel trial) and even a new word added to the *Merriam-Webster's Collegiate Dictionary* in 2003 (McJob).

Multiple names for the same thing make understanding more difficult. This is why referential disambiguation is needed to parse "them," "they," "it" and an infinite variety of slang and colloquialisms. Humans excel at this activity, but without enough human readers, computers must be employed to interpret, identify and tabulate this new form of brand experience.

Once the number of exposures has been tallied for a given brand or product, the total must be segmented by subject ("I love this product" vs. "It doesn't work") and then by sentiment ("It doesn't work" vs. "It doesn't work, and I hate it"). This is where text analysis in social media becomes critical for customer service and customer care.

### Trouble Spotting

With ever-growing piles of survey responses, customer service representative comments and online feedback, customer service departments were the first to recognize the value of text analytics. Every customer service employee has an opinion about the most frequently mentioned problems and their relative severity based on frequency and customer tone of voice. But only a comprehensive analysis can validate those feelings.

Ford may have discovered their Explorer/Firestone tire problem faster, and Toyota may have been able to move quicker on their accelerator problem if they had analyzed trends in their customer service databases.

---

■ Public opinion is trusted more than advertising.

---

In 1994, Intel might not have been the target of a CNN story if they kept an eye on online discussions about their products. Intel tried to minimize the floating-point division problem in their Pentium Pro chip, suggesting it wouldn't have an impact on most users. They offered to replace the chip if the customer could prove they had been affected. The online outcry became a mass media spectacle and Intel learned to be more attuned to this new form of public conversation.

These incidents can have a cumulative effect on the overall equity of a brand - something that companies are well-advised to track.

## Brand Affection Monitoring

A company used to be able to create a brand image in the minds of the public through advertising alone. Today, advertising is only a part of the brand-building process. Consumers are making more deliberate, considered purchases as opposed to impulse purchases, which are all researched online. That research inevitably includes reviews by experts with blogs, customers with opinions and friends with recommendations.

The quantity and polarity of online, written communication about almost any product or service has become statistically significant and sways purchases. Given the trust imbued in friends and experienced customers and the skepticism felt toward promises delivered by vendors, what people say about you online is becoming more influential than advertising.

If a soap company claims their new formula makes skin softer and three influential bloggers post to the contrary, it will only take days before the soap company's television ads become objects of satire on YouTube, which could go viral.

If a promotional experiment, like the Old Spice Guy, goes viral, only social media monitoring tools can help determine if the millions of views are generating positive feelings or are making Old Spice the butt of a million jokes.

The health of a brand is no longer measured in simple sales - that's too late. An organization must stay on top of public opinion to respond promptly and properly to a viral gold mine or a public embarrassment. The days of random-digit-dialing phone calls for surveys are over and the days of listening to what people have to say about you, rather than to you, are upon us.

---

■ Tracking subject matter and polarity acts as an early-warning system.

---

## Identifying Opinion Leaders

Every industry has its throng of influencers and finding them is no longer a matter of asking a high-priced PR firm for a list. Identifying and influencing the influencers is a new skill, much different from wining and dining journalists and industry analysts.

An 11-year-old Chicago girl named Tavi Gevinson became the center of attention in the fashion world because of her blog, Style Rookie. Her opinionated prose is fresh and blunt and she is quite talented. Now 14, Gevinson is in demand at Paris fashion shows, on the speaking circuit and was the topic of an eight-page spread in *The New Yorker* (Sept. 20, 2010).

Those in the fashion industry monitoring what is being said by whom and how much they are repeated and retweeted - and by whom - have been aware of Gevinson for years. They know this young lady is a force to be watched, enticed and befriended. The rest have been taken by surprise.

But simply listening and evaluating the chatter is only the beginning. It's time to join in.

## Monitoring One's Impact on the Conversation

While the conversation in the marketplace careens from subject to subject and oscillates from positive to negative, every organization does its best to participate. The first goal of establishing an active social media voice is to simply be a visible member of the community.

Actively participating in the public dialogue can make a firm look tuned-in or clueless, caring or arrogant, accommodating or antagonistic. Knowing how the sociosphere is responding to corporate efforts is crucial.

The second goal is to actually influence the discussion and have an impact on brand sentiment. Controlling the message is no longer possible. Managing the message is everyone's goal. The reality is that an organization can only hope to influence the conversation and must be diligent about monitoring its own success.

Text analytics come to the aid of marketers trying to understand the best ways to participate. Observing the response to online corporate socializing is akin to gauging a dinner party's response to your part of the conversation. If you do not contribute, you won't be invited back. If you are overbearing, you will be shunned. It's called social media for a reason.

As always, a review of the numbers is interesting and tells you what happened. But you can use text analytics to look into the future.

---

■ Sales metrics are the rearview mirror of marketing efforts.

---

## Trend Analysis and Predictive Modeling

Tracking the number of viewers, followers or commenters is interesting, but only the amount of change in that number is useful. The more you track public opinion over time, the more likely you are to recognize trends.

The excitement at the launch of a new game may be replaced with frustration as people try to learn how to play, followed by elation at learning hidden tricks and the long slide to complacency as the thrill wears off.

Attitudes about your brand or products may change seasonally. Response to your offering may be affected by many factors, including weather, sports seasons, holidays or fiscal calendar.

Monitoring sentiment over time not only provides a map of the popularity of your brand over time, but can offer clues as to what to expect for the next interval.

We can now plot promotional activities against a historic map of ongoing consumer response. This lets us logically create compelling social media campaigns and better understand the results.

---

■ It's not what you know, but whom you retweet.

---

## V. Text Analytics in the Future

Let's step from predictive modeling to predicting the uses of text analytics in social media. This area is changing rapidly, making it hard to say if the following scenarios are predictions or merely logical conclusions.

### Twitter Reveals Moviegoers Partiality

"Ross, why are you changing out the promos?" Peter was nervous, on the edge of panic. "We cut *Preview Series A* to show more car chases. The focus groups went crazy for the car chases when we tested the movie on them. That's our audience!"

Ross shook his head. "It was our audience. Not anymore."

As a movie promoter, ticket sales and critics reviews were the only metrics Ross could depend on. He chooses the best preview to distribute based on focus group responses but that was a hit-or-miss proposition. That was then.

"Whaddaya mean, 'Not anymore'? When did you get a crystal ball?"

"Since I read this paper from HP Labs." Ross tossed a copy at Peter. *Predicting the Future With Social Media* ([http://arxiv.org/PS\\_cache/arxiv/pdf/1003/1003.5699v1.pdf](http://arxiv.org/PS_cache/arxiv/pdf/1003/1003.5699v1.pdf))

“OK, so you’re a scholar. I’m duly impressed now boil it down for me.”

“Just read the abstract at the top.”

Exasperated, Peter grabbed the paper from Ross and read:

In recent years, social media has become ubiquitous and important for social networking and content sharing. And yet, the content that is generated from these websites remains largely untapped. In this paper, we demonstrate how social media content can be used to predict real-world outcomes. In particular, we use the chatter from Twitter.com to forecast box-office revenues for movies. We show that a simple model built from the rate at which tweets are created about particular topics can outperform market-based predictors. We further demonstrate how sentiments extracted from Twitter can be further utilized to improve the forecasting power of social media.

“So, Mr. Twitter Scientist, you’re changing our advertising because of something a little bird told you?”

“Yeah - but it’s about 25,000 little birds. Real, live audience members are tweeting spontaneous reviews and I’m adjusting our promotions accordingly. If people are pleasantly surprised that the movie contains lots of car chases, I can see to it that previews with more chase scenes get on TV and into theatres. But that’s not the case.

“Women are tweeting about our leading man’s dreamy eyes and guys are tweeting that it’s a great date movie,” Ross explains. “So we’re pushing the car chases to the back and bringing the love story forward ... *this week*.”

“This week?”

“Things change. It pays to watch the birdie ...”

## Social Media Frightens Fast Food

Executives at a fast food chain are blindsided by reports of an outbreak of food poisoning at their restaurants. Their internal processes for problem reporting did not surface any issues so the management team is jolted into a frenzy of activity.

Emergency plans are initiated to launch an investigation of all food suppliers and preparation processes. Layers of management are to be reassessed to determine where the chain of command fell apart. Consultants are to be brought in for independent audits. The public relations department and agencies are to be energized to develop corporate communications response proposals for board review while keeping the press at bay.

But, all that can be provided today is a press announcement saying the company is aware of the problem and is looking into it.

At home that evening, the chief marketing officer's daughter pats her on the back and says, "Sorry you're having such a tough day, Mom. Twitter-vicious, huh?"

Before making another move, the CMO opens the e-mail from her text analytics team that had been waiting for three days.

Hi Barbara -

We've noted a spike in negative comments about our french fries. If this gets picked up by the news networks it could get twisted, so we wanted to give you a heads up.

The fries are fine. The problem started on a public answer board Q&A that suggested vegans and vegetarians could get sick if they eat fries that contain natural flavorings derived from meat products or were fried in oil that meat products are cooked in.

This is an old issue - something we addressed years ago, but for some reason, some dolt thought it was worth ranting about.

It got reposted and retweeted and now it's "a thing."

This is one of those times where a nonresponse is best. In this environment, it's important to not fan the flames, just let it burn itself out. If it becomes national news, then we can discuss the best ways to nullify its effects. We don't want this turning into a Nestle/palm oil/guerilla killer-type situation.

Sally

Barbara looked at the press announcement and realized the company had done exactly the wrong thing. She e-mailed Sally and the social media marketing department asking,

"Now that we've put our foot in our mouth, how do we keep from shooting ourselves in the foot?"

The response was short and sweet:

"Ask the CEO to issue the following tweet:

Switched fries 2 veg oil 3 yrs ago. Tasty and safe! Try 'em on me this Wednesday - Say 'Meatless Fries' w/ Healthy Meal order. Plz RT."

Not only was the crisis averted, the press picked up the story as a successful how-to example, the CMO got an increase in her social marketing budget and the text analytics team e-mails got priority attention in the future.

## **Social Media Taps Hearts and Minds**

When an auto manufacturer rolls out a concept vehicle at a car show, there's a bit of showing off involved. But mostly, it wants to know if its concept is palatable to the buying public. In the past, it could get a response from a couple of dozen people and hope they were representative of others. Today, it rolls out a new concept car and listens to online conversations to see how people describe new features, whether they grasp the benefits and if they have any suggestions.

A new B&B opens and immediately solicits guests to post comments about their stay. Not only does it get more publicity, it gets to see how people describe their experience. These descriptions help it to tailor the description on its website and in its promotions. It also gets to quote "free-range" testimonials. In addition, monitoring this indirect feedback lets it stay on top of common complaints, redress grievances and identify areas for improvement in its offering.

In their desire to hear directly from customers, some companies are setting up in-house listening websites and inviting people to offer ideas. This digital suggestion box offers participants the chance to vote on submissions. As a result, companies like Starbucks (<http://www.mystarbucksidea.com>) and Dell (<http://www.ideastorm.com/>) collect thousands of opinions and can instantly see which are the most popular. Using text analytics, they can see which suggestions are the most emotionally charged, show up the most often or come from specific regions or customer segments.

Casting a wider net, many companies use social media text analytics to find new niche blogs and websites that would be a match for their advertising message. They can uncover affinity groups and clusters of people who are likely to be more interested in their goods and services. A horse shampoo manufacturer that has been marketing to horse breeders and veterinarians discovers a subset of beauty consultants lauding the benefits of their products for humans. A glass company is surprised to find architects discussing a competitor's product as a structural element in their buildings. Insurance companies find that tweets and posts tell a fuller story of local floods and fires for risk management. A dishwashing detergent is lauded by volunteers and professionals cleaning up oil-doused birds.

Listening to people talking about you is much more revealing than analyzing what they say to you.

---

## VI. Conclusion

---

With social media as firmly in place as the Internet itself, monitoring the ebb and flow of public attitude is now a business necessity. Keeping a finger on the pulse of public opinion is now a part of marketing, from advertising and public relations to customer relationship management and customer lifetime value.

Statistically derived intelligence from business-rules-driven text analytics tools is the latest advantage in the ongoing struggle for relevance, timeliness and a competitive edge.

Coupled with alert functions and workflow capabilities, these monitoring and sentiment analysis tools provide modern marketing professionals with a way to judge the impact of their promotional expenditures; allow large organizations to respond in a more personal manner to an ever-changing marketplace; and demonstrate to a more skeptical public that the company is less mechanistic and more personable than ever.





**SAS Institute Inc. World Headquarters +1 919 677 8000**

To contact your local SAS office, please visit: **[www.sas.com/offices](http://www.sas.com/offices)**

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. Copyright © 2010, SAS Institute Inc. All rights reserved. 104806\_S63591.1110