LEXALYTICS
Read Between The Lines

Angoss
Predict. Act. Perform.™

WHITEPAPER

Text Analytics - Sentiment Extraction

Measuring the Emotional Tone of Content

What is Sentiment Scoring—and why do you care?

Sentiment scoring allows a computer to consistently rate the positive or negative assertions that are associated with a document or entity. The scoring of sentiment (sometimes referred to as tone) from a document is a problem that was originally raised in the context of marketing and business intelligence, where being able to measure the public's reaction to a new marketing campaign (or a corporate scandal) can have a measurable financial impact on your business.

Historically, the measurement of tone was the responsibility of the marketing department in an organization and was typically done by hand. The obvious limitations of scoring content by hand led to the development of machine scoring.

Once you have reliable, consistent machine-based sentiment scoring, there are a number of applications that become feasible inside of financial services (automated trading, better information to traders), reputation management (the problem every marketing person faces), "voice of customer" (listen to how they're saying what they say, don't constrain them to closed-ended questions) and eDiscovery (was there a wave of negative emails before a certain crisis hit?), among others.

This paper will demystify sentiment scoring and explain how the Lexalytics sentiment engine works. This includes a discussion of how and why the basic concept of document sentiment has been extended to the paragraph and entity level, and how this technology is being further extended to measure other indicators within content, including the assessment of threat, customer satisfaction and many other contextual indicators.

How does Sentiment Scoring work?

Humans seemingly have no trouble reading a sentence and mentally scoring the sentiment. We humans use a process of reading and understanding the descriptors placed on the subject of a sentence.

Consider these sentences:

- A horrible pitching performance resulted in another devastating loss.
- Sub-par pitching and superb hitting combined to cost us another close game.

They both have the same basic subject, the loss of a baseball game, but obviously (to you!) the first sentence is contextually much more negative. The keys that humans use to discern this are to focus on the emotive phrases "horrible pitching" and "devastating loss".

The sentiment system developed by Lexalytics does exactly the same thing.

The Lexalytics engine identifies the emotive phrases within a document and then scores these phrases (roughly -1 to +1), and then combines them to discern the overall sentiment of the sentence.

Importantly, the sentiment scoring will score those sentences the same every time they're exposed to them – the engine is not affected by whether or not it has had its morning coffee, or whether it is a fan of the team that lost (or that won!). This consistency is very important.

Part Of Speech (PoS) Tagging

The first step in determining the tone of a document is to break the document into its basic parts of speech (POS tagging). POS tagging is a mature technology that identifies all the structural elements of a document or sentence, including verbs, nouns, adjectives, adverbs, etc.

Even though you don't really realize you're doing it, to determine the sentiment of a document, you're mentally identifying the parts of speech within a document that indicate emotion. In most cases these are adjective-noun combinations like "horrible pitching" and "devastating loss". *That's how the Lexalytics engine works, too.* These combinations are called "sentiment-bearing phrases"

The difference, of course, is that the text analytics engine needs to actually assign a number to the sentiment – as opposed to you, when you're reading it, just think "darn, they lost again…badly". What the engine does is create a very, very large dictionary of sentiment bearing phrases and their relative scores. These scores are pre-determined by how frequently a given phrase occurs near a set of known good words (e.g. good, wonderful and spectacular) and a set of bad words (e.g. bad, horrible and awful).

We used an extremely large corpus of text (the Web, via an internet search engine) to evaluate the nearness of known good and bad words to the phrase being considered. Consider the case of the phrase "devastating loss". It means something to you because you've come to associate that with something bad. That's exactly the process that we go through – we check to see if we should associate that phrase with positive or negative sentiment, and just how closely we should associate it.

Thus we use search engine queries as so:

"(devastating loss) near (good wonderful, spectacular)"
"(devastating loss) near (bad, horrible, awful)"

Each query comes back with a hit count. These hit counts are combined using a mathematical operation called a "log odds ratio" to determine the score for a given phrase. In this case the phrase "devastating loss" yields a phrase score of negative 0.56.

Here's how this would work in a document:

Dan Wells joins Nautilus Footwear, America's fastest growing safety shoe company, as Vice President Nursing Division. "We have ended 2000 with record sales and earnings. Phase 3 of our strategic plan is to clearly separate our three business units of industrial, nursing and public safety. Dan will lead our Nursing strategy and drive the efforts of our nursing sales team. The clear separation of our Nursing business is a natural for us as we have been a leader in the nursing arena since beginning the company in 1996," said Wayne Easley, President / CEO. "Dan comes to us with an extensive background in the footwear industry most recently with Berkshire Hathaway's, Lowell Shoe Company. He has solid account and product knowledge that can clearly continue to deliver our 'ergonomic message' to the Nursing community," added Elsey.

The green phrases are the sentiment bearing phrases in the document (these all happen to be positive sentiment bearing phrases):

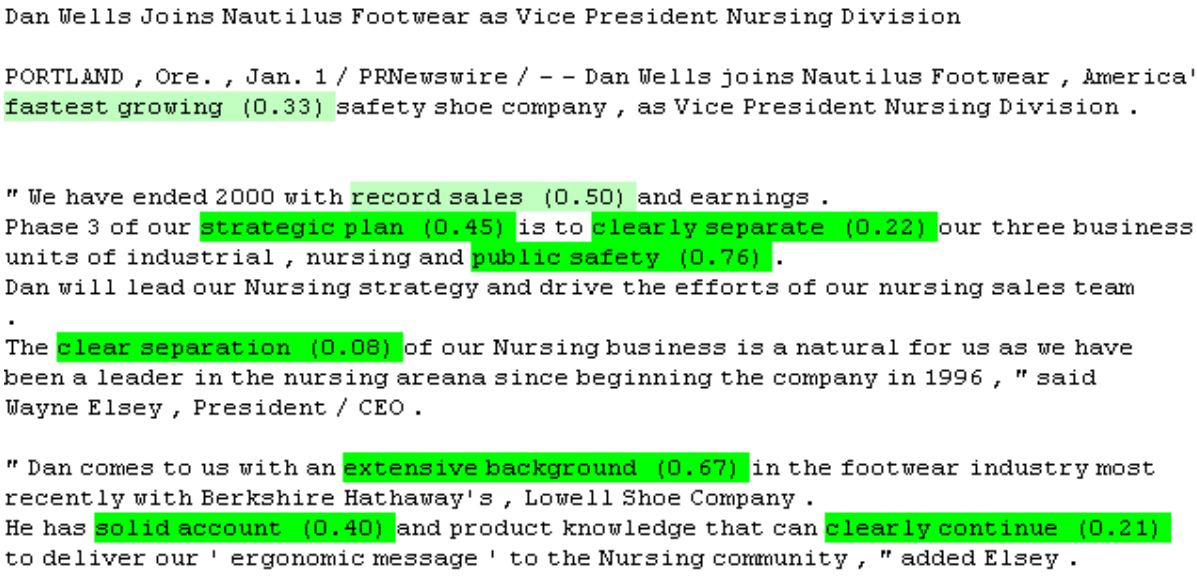

Figure 1

Whole Document Sentiment Scoring

We have an algorithm that we use to combine the phrase scores in the document based on an operation called "lexical chaining". This operation is beyond the scope of this document, but is similarly consistent and repeatable. The overall document sentiment for this document comes out to 0.365 (mildly positive).

Going Beyond Document Sentiment

Ok, you know how we do it for a whole document. Unfortunately, except for press releases, it's rare to find a document that is completely homogenously positive or negative. Sentiment is typically a localized phenomenon that is more appropriately computed at the paragraph, sentence or entity level.

Consider the following example:

"Julie Jones' superb performance in the gubernatorial debate has all but assured her of victory in the upcoming elections. Unfortunately, the evening did not go as well for her opponent John Adams. Mr. Adams' nervous and uncertain performance has put his entire political future into question."

The sentiment of this sentence is completely different for the two individuals described within, while the overall sentiment for the sentence averages out to roughly neutral.

Let's look at the results of this snippet at the overall level and at the entity level (red are negative sentiment bearing phrases, green are positive):

Sentiment Tagged Text:

Julie Jones superb performance in the gubernatorial debate has all but assured her of a majorvictory in the upcoming elections. Unfortunately, the evening did not go as well for her opponent John Adams, his nervous and uncertain performance has all but guaranteed a loss and put his entire political future into question.

Figure 2 Sentiment Tagged Text

Now let's examine the entity level sentiments for each person, and start by identifying every instance where the person is mentioned. Again, how we determine just what constitutes an entity is somewhat beyond this document, but isn't important for understanding how we assign sentiment:

Entity Tagged Text:

Julie Jones superb performance in the gubernatorial debate has all but assured her of a major victory in the upcoming elections. Unfortunately, the evening did not go as well for her opponent John Adams, his nervous and uncertain performance has all but guaranteed a loss and put his entire political future into question.

Figure 1 Entity Tagged Text

In the above-tagged text, notice that the system identifies not only the people by name, but also identifies the pronouns “her” and “his” (and will correctly associate the “her” with Julie Jones – an operation called “pronominal co-referencing”). Correctly including the pronouns significantly improves our software’s ability to measure the tone for each individual. Running this block through and computing sentiment for each entity yields:

- Julie Jones: Positive (+)0.22
- John Adams: Negative (-)0.11

The tagged text yields a document sentiment of 0.11 which is squarely in the middle of the neutral range, and doesn’t really tell us anything about the real tone of this snippet.

This capability sets our core software apart from other products that measure sentiment, and enables our software to focus on the sentiment or tone of specific people, companies or products. The true value of measuring sentiment is in applying the measure to the people or products you’re concerned about. For most rich content sources, it’s much more important to measure and compare the sentiment of an individual entity than it is to get the overall sentiment of the document. Consider the case of a product review article, where one product gets panned and the other doesn’t – if you can’t discern which was which, then it doesn’t do you any good.

Summary

The determination of Sentiment or another indicator is another step in the process of converting unstructured content to structured content, so that the humans who interact with this ever-expanding sea of information can spot trends and patterns within the content.

The last few years have seen significant enhancements in the creation of metadata from content, including the extraction of entities (People, Places, Companies, and Products), key noun phrases, and subject classification. The Lexalytics Salience Engine has 7 years of experience in teaching computers how to consistently and reliably measure emotion, and more importantly, how to appropriately ascertain just to whom in the article that sentiment is directed.

About Angoss Software

As a global leader in predictive analytics, Angoss helps businesses increase sales and profitability, and reduce risk. Angoss helps businesses discover valuable insight and intelligence from their data while providing clear and detailed recommendations on the best and most profitable opportunities to pursue to improve sales, marketing and risk performance.

Our suite of desktop, client-server and in-database software products and Cloud solutions make predictive analytics accessible and easy to use for technical and business users. Many of the world's leading organizations use Angoss software products and solutions to grow revenue, increase sales productivity and improve marketing effectiveness while reducing risk an cost.

About Lexalytics, Inc.

Lexalytics, Inc. is a software and services company specializing in text and sentiment analysis for social media monitoring, reputation management and entity-level text and sentiment analysis. By enabling organizations to make sense of the vast content repositories on sources like Twitter, blogs, forums, web sites and in-house documents, Lexalytics provides the context necessary for informed critical business decisions. Serving a range of Fortune 500 companies across a wide spectrum, Lexalytics partners with industry leaders such as Endeca, ThomsonReuters, Radian 6 and TripAdvisor to deliver the most effective sentiment and text analysis solutions in the industry.

Corporate Headquarters

111 George Street, Suite 200
Toronto, Ontario M5A 2N4
Canada
Tel: 416-593-1122
Fax: 416-593-5077

European Headquarters

Surrey Technology Centre
40 Occam Road
The Surrey Research Park
Guildford, Surrey GU2 7YG
Tel: +44 (0) 1483-685-770

www.angoss.com