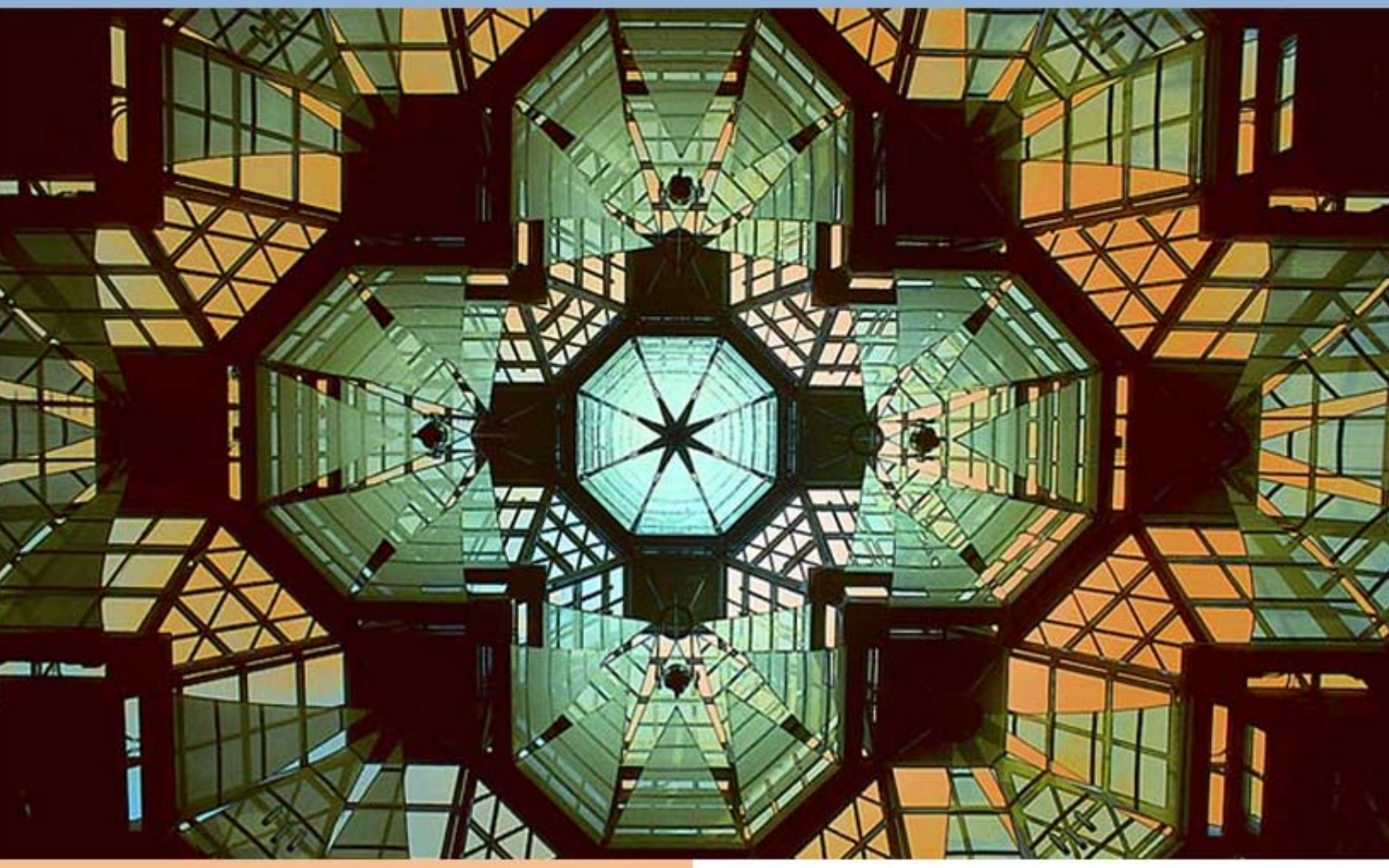
WHITE PAPER

FINDING REAL INTENT IN BIG DATA

HOW TO GET HIGHLY ACTIONABLE INSIGHTS FROM TEXT ANALYTICS



neurolingo.com

NEUROLINGO, L.P.

17 April 2012

SUMMARY

The buzz today is all about "big data" – harvesting information and insights from large-scale data sets found via the Web, social media and enterprise servers. There are lots of text analytics tools today that can pick out key word patterns from large data sets to give you statistical insights into what's being said in structured and unstructured data sources. But Neurolingo is ahead of the crowd in its ability to identify and to apply the rules of language structure to any type of large data set to find not just words but the focus of what's being said and the intent that can be inferred from it.

Most importantly, the highly adaptive and rapidly programmable design of Neurolingo technologies can be used to develop services focused on highly specialized types of content and language, such as specific professions, topics, localities, interests and special grammars specific to communities and publishing platforms. This can allow you to develop broad-scale applications that can apply common analysis and logic to many different types of information sources with different types of grammar and language structure easily, delivering insights that can span many types of communities or that focus in on specific communities.

*Are you trying to find not just sentiment or key concepts in big data but also the **focus and intent** of very specific types of people for very specific topics and purposes?*

This case study shows how Neurolingo's advanced text analytics platform delivers measurable and remarkably accurate forecasts from unstructured social media data as a key example of how you can turn content from any big data source into insight.

This paper outlines why being able to extract real focus and intent from big data is important, how Neurolingo does it and offers a key example from Neurolingo's highly successful application of this technology to interpreting social media sources to forecast the outcome of major sporting events. The lessons from this example can be applied to any enterprise or media operation to deliver highly valuable insights from vast amounts of data from unstructured text sources.

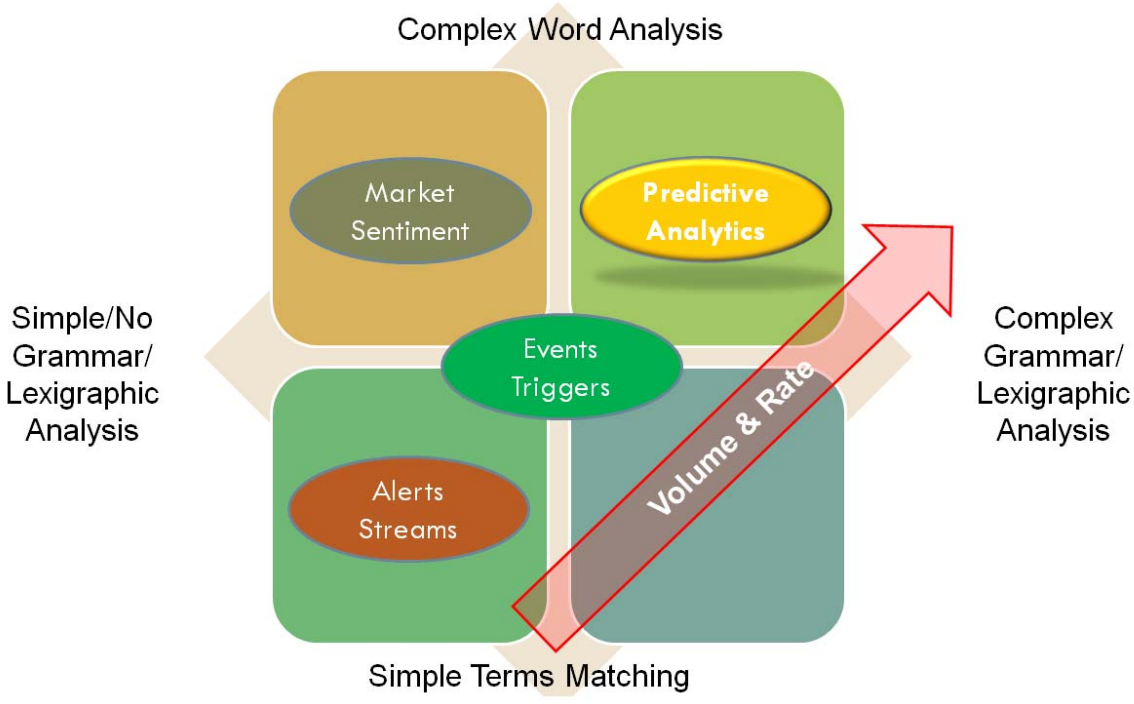
HOW DO YOU FIND VALUE IN BIG DATA?

Why are some people focusing on "big data" so much? Well, because it's big. According to CenturyLink Business estimates, in 2011 the world generated 1.8 zettabytes of new and copied data – that's almost two thousand billion, billion bytes of data in one year. By 2015, that rate is expected to climb to 7.9 zettabytes – the equivalent of 18 million times the contents of the U.S. Library of Congress! And while personal computers contribute quite a bit of this new data, data from mobile devices will grow about 82 percent by 2015 and data from non-PC devices about 37 percent. Put simply, there's more data from more places than ever before requiring analysis by people who need to make important decisions of all kinds. Much of this new data is "unstructured" – information in loosely organized formats such as word processing documents, ebooks, text messages and video transcripts. So not only is there an enormous amount of new data, the "hooks" to process it into a meaningful form are often lacking.

Harvesting insights from this deluge of new information requires finding patterns in these huge and rapidly changing unstructured data sets fast enough and accurately enough that they can actually make a difference in decision making processes. However, doing this requires not just powerful technology, but also knowing where to focus your efforts with that technology. You need to find both meaningful structure in these enormous information sources but also the human relevance that can be inferred from that structure in a meaningful time frame.

For example, during February of 2012 hundreds of millions of people in the U.S. and around the world were watching the National Football League Super Bowl XLVI championship game. During the game, sports fans were using the Twitter social media service to broadcast an average 12,233 text messages per second to their followers – that's a social media source for a single sporting event generating almost ten percent of the record rate for stock trading messages on the New York Stock Exchange. But which of those messages really said something about how people thought that the game would turn out in time to affect any decisions that people were making about it? It's pretty challenging to extract that type of concrete insight from a sea of text messages.

For any "big data" challenge - be it in sports, finance, health care, logistics, government or consumer marketing – you need to understand that not all technologies are equally equipped to deliver meaningful insights from huge sources of unstructured text information. Note in the chart below that when it comes to analyzing text using software, depending on what you're trying to understand in large data different tools with different type of analysis techniques may be appropriate.



Relation of text analysis tool sophistication to support of specific analysis goals

As you can see, as the volume and rate of information increases, it becomes more practical and more necessary to apply more sophisticated text analytics tools, which can support more complex analysis goals. The most valuable type of text analysis is predictive analytics – being able to forecast likely behaviors of people or expected outcomes for specific types of events or processes based on the analysis of documents and text from data sources such as social media and emails. Predictive analytics software has been around for many years, but its was focused mostly on number crunching based on structured data harvested from large databases on enterprise computer systems. To tame the flood of "big data", predictive analytics now has to extract meaning from unstructured data – and it's not a simple task. It requires very sophisticated text analytics software to do it right.

*For successful text analytics that can forecast likely events and actions, you need sophisticated software that can determine the **focus** and **intent** of an information source for a wide variety of **topics, cohorts** and **technology platforms**.*

There are two key things that any text analytics must be able to do to support predictive analytics effectively:

- **Understand the *focus* of your information.** Whatever text you happen to be processing, being able to understand the focus of a particular message or document is critical to finding the "needle in a haystack" that matters to you. Less sophisticated text analytics tools do this by matching words and word patterns to come up with topics that a document may be focusing on. But often there are details in text information which can determine its focus that can only be attained by "reading" the document – getting down to the word structure and grammar of a particular source. In some sources of information, word structure and grammar is relatively easy to understand, such as in newspaper articles or other well-edited sources.

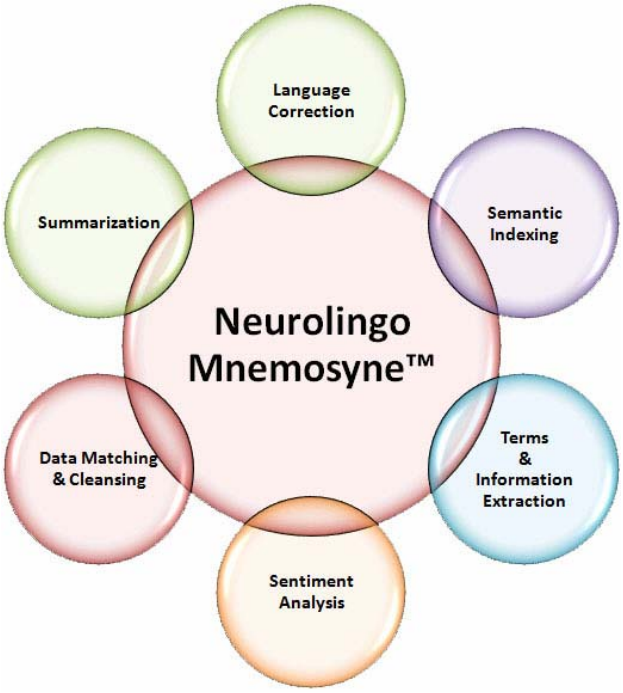
  But in information sources like social media text messages, where there is "grammar" that makes sense in short messages that wouldn't make sense in a news article, this can be much harder. The same problem may be found with informal language and grammar used in sources that might be very specific to an industry, a location or even a specific topic. So the more unstructured and unrefined the source of text, the harder it is for text analytics software to find a meaningful focus to its information.

- **Understand the *intent* of your information.** If you're able to find the focus of a document or message, the next question is whether you can infer people's intent from that source to build predictive models of likely events or behaviors. In other words, now that you know what someone is really talking about, what might they be thinking of doing or what might they think is going to happen? This is a critical type of insight that can only be found by looking very carefully at the detailed structure and grammar in a source of text. And it's a lot more than typical sentiment analysis text analytics software can handle, since they focus largely on counting the number of times specific words or concepts show up in text sources.

Yet again, having software that falls short of sophisticated and flexible analysis of unstructured text sources can leave your analytics out in the cold. And it gets harder when you try to encompass more different types of analysis and information sources. Then this type of analysis has to be tuned to a wide variety of topics, to specific groups of people (cohorts) and the way that they communicate, and to any number of ways that people communicate due to the type of software and hardware that they use to send and receive information. Each one of these – **topics**, **cohorts** and **technology platforms** – introduces unique and ever-changing wrinkles to defining and interpreting the grammar of text to a degree that can deliver meaningful insights into the intent of an information source.

THE NEUROLINGO SOLUTION

The challenges to delivering valuable predictive analytics and other sophisticated text analytics solutions from huge, unstructured text sources are undoubtedly significant. There are few who have been able to address these challenges effectively for the development of information services and solutions tailored to the needs of a wide variety of enterprises. One of the leaders in these cutting-edge text analytics solutions is **Neurolingo**, L.P., which since 2005 has developed a sophisticated and flexible text analytics technology platform that has helped major enterprises to solve very challenging text analytics problems. Neurolingo has packaged its XML-based text analytics technologies into a readily deployed and tailored solution called **Mnemosyne™** - named after the goddess of Greek mythology responsible for the nine muses of human creativity and the power of human memory.



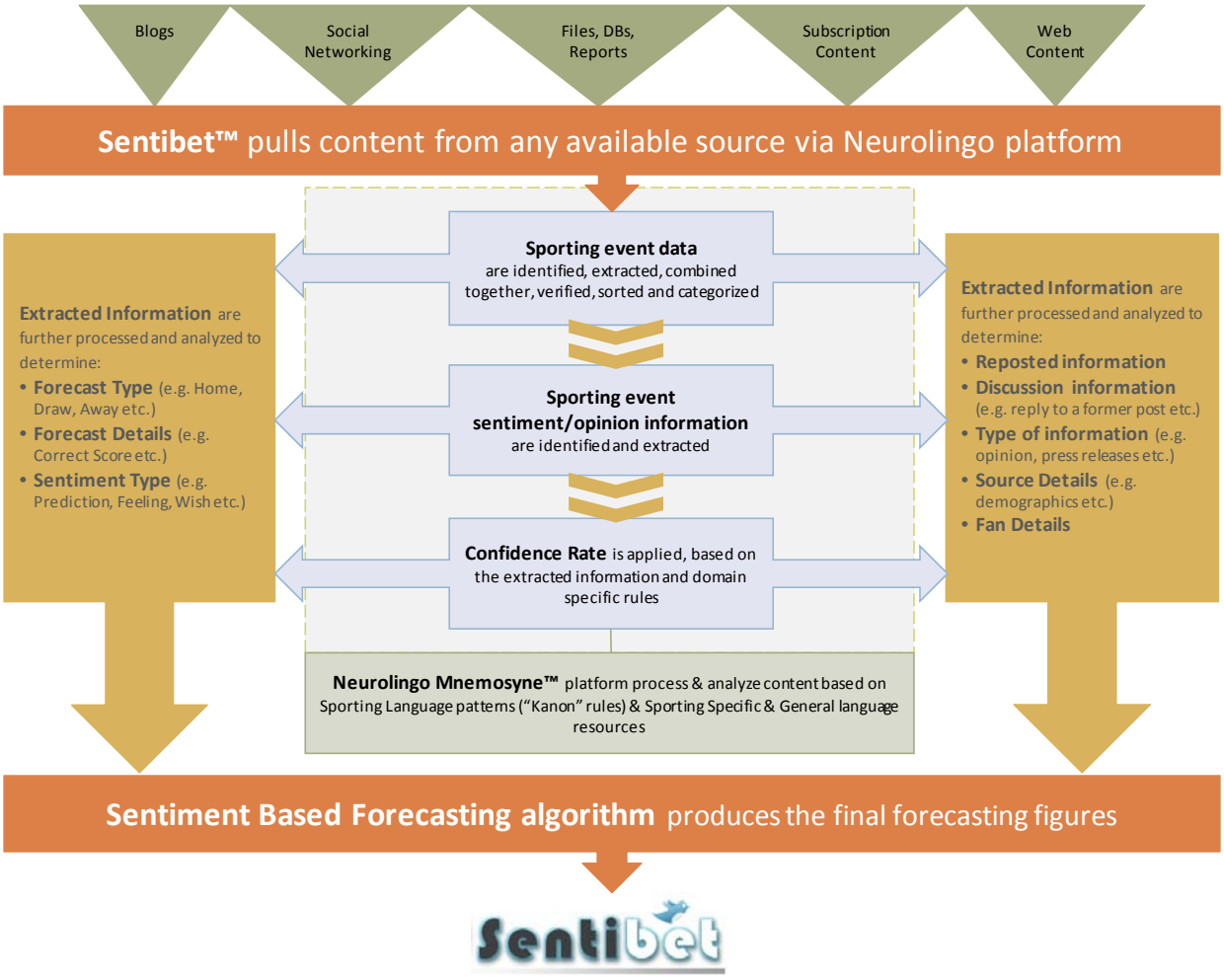
Like the goddess of Greek mythology, Neurolingo's Mnemosyne technology has many talents and can help us to "remember" what was lost in seas of information that overwhelm us. Mnemosyne combines powerful word analysis, text and word structural analysis and grammatical and semantic analysis of any text source in a highly flexible and easily programmed software system. This means that Mnemosyne can be tuned to look at any information source or any combination of topics, cohorts and platforms found in any number of information sources as a unique system of language with its own rules and structure. Since the rules, vocabularies and semantic analysis of Mnemosyne can be programmed with this level of ease and flexibility, it can be constantly tuned and re-tuned to keep up with dynamic changes in information sources. This enables Mnemosyne to continually improve its ability to identify the focus and intent that can be derived from any information source. Mnemosyne can be deployed as

an installed software solution for private use or as a Web-based "cloud" service that can digest information and deliver analysis.

CASE STUDY – SENTIBET.COM

The "how" of Neurolingo's Mnemosyne technologies is impressive, but more impressive is what it does when applied to very challenging predictive analytics problems in "big data" sources. As a demonstration of Mnemosyne's capabilities, Neurolingo developed the **Sentibet.com** service, which provides forecasting of sports game outcomes based on its analysis of messages on the Twitter social media service in conjunction with supporting information sources. Neurolingo focused Sentibet.com's processing on major sporting events that generate significant streams of Twitter messages that enabled the Mnemosyne technology to be tuned to the various styles of semantic information found in Twitter messages specific to teams, geographies, events and the peculiar styles that each of these generate in Twitter messages that people "tweet" from their PCs, mobile phones and tablet computers.

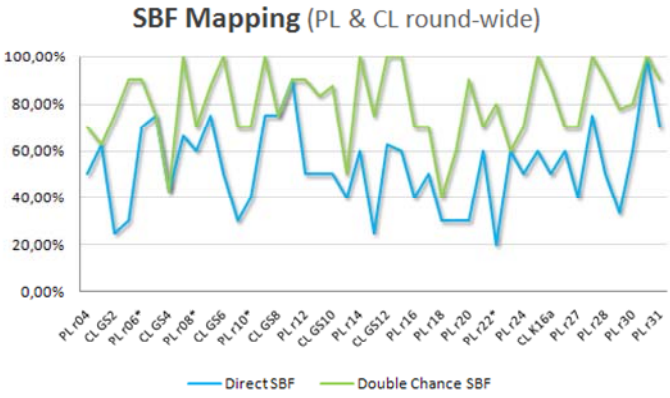
The Sentibet.com project put all of Mnemosyne's text analytics capabilities to a very full test:



Sentibet.com Processing Flow via Neuroingo Mnemosyne Platform

As can be seen in the above diagram, Sentibet was tuned to identify first when a Twitter message was related to a particular sporting event, then to process what kind of focus and intent was found in a message related to forecasting a game's outcome. There are three main areas of forecasting that Sentibet processes – whether a message is conveying a **feeling** about a game's likely outcome, a **wish** for its outcome, and a **prediction** of its likely outcome. These three semantic concepts were then analysed using a variety of specific formulas to come up with an overall prediction model for the percentage of likelihood of an outcome for a game based on people's forecasting information in their Twitter messages.

The results were displayed in the Sentibet.com portal, including both the overall graphs of analysis results and a stream of Twitter messages with indications of how Sentibet interpreted the message – was it a message indicating a feeling, a wish or a prediction and what type of outcome they expected. People visiting the portal could suggest corrections to the Sentibet analysis of specific messages, enabling a feedback loop that helped the Neurolingo team to refine its results. As it continued to accrue forecasts over dozens of games in the Premier League and Championship League for European football and U.S. games in the National Football League and NCAA college basketball games, Sentibet.com was also able to add factors based on the ability of specific contributors to forecast outcomes – "expert opinions" of a sort.



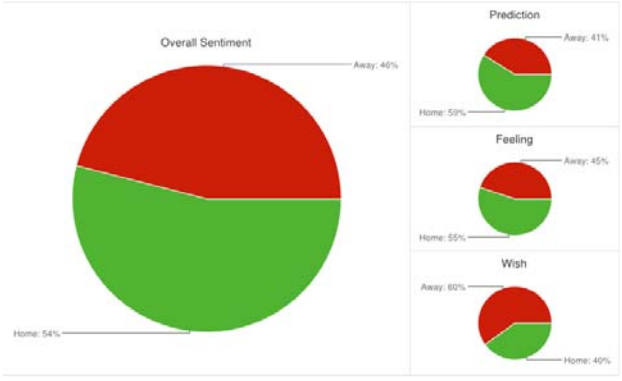
The results of this demonstration have been remarkable, both for Sentibet's ability to extract messages with the correct focus on specific games but also in its ability to correctly identify the forecasting intent of those messages. Over several months of Premier League and Championship league matches, Sentibet was able to forecast the "double chance" of either a win or a tie for specific matches in specific rounds of competitions **60 to 100 percent for almost every round** – consistent with or well ahead of many odds-makers' ability to provide accurate forecasts.

Also remarkable was Sentibet's performance in once-a-year championship matches in U.S. sports. For the 2012 NFL Super Bowl, Sentibet identified more than 87,000 Twitter messages in the days leading up to Super Bowl XLVI relating to forecasting the outcome of this high-profile contest. The analysis required rapid re-tuning of Sentibet's Mnemosyne text processing tools to accommodate some significant

differences between the semantic content of Twitter messages relating to the teams and the different aspects of American sports and fan terminology.

Although many odd-makers were picking the New England Patriots to win this game, the Sentibet analysis was showing a consistent fan forecast for a New York Giants win. During the game itself, the millions of messages being sent during the match revealed more than 5,000 additional messages relating to forecasting its outcome – messages that continued to support a Giants win. The ability of Sentibet to find not just messages with sentiment but messages with both focus and intent specific to forecasting game outcomes seems to have made the difference. It was a close win by the New York Giants in overtime, which was also reflected by the relatively tight percentage of forecasting fans putting the Giants as the eventual winners.

Later in 2012, Sentibet turned its focus in U.S sports to the NCAA college "March Madness" basketball tournament. With much more focused followings, college basketball games generates smaller numbers of messages on Twitter providing forecasting information. For the final match of this tournament, though, Sentibet was able to identify more than 3,400 Twitter messages related to the match's outcome – a statistically significant sample of opinions. A strong majority of fans expressing a forecast on Twitter via Sentibet analysis were wishing that the Kansas Jayhawks would win the tournament – and so were many prominent sports odds-makers. However, both their feelings about a likely outcome and their predictions were forecasting a win by the Kentucky Wildcats. Kentucky went on to win the NCAA championship college basketball game.



The Sentibet platform developed by Neurolingo seems to demonstrate conclusively a few key factors relating to developing sophisticated semantic analysis based on unstructured data:

- **Finding real intent in big data can work** – if you apply the right sophisticated tools to the right problem to identify the correct signals with the correct analysis.
- **Highly accurate semantic extraction can maximize results** – you can get statistically significant results from relatively small amounts of text if you have highly accurate identification of its structure and meaning.

- **Gauging sentiment is not enough** – you have to get into the real language of a source to get at the real focus and intent of specific cohorts using specific types of language.

HOW NEUROLINGO CAN WORK FOR YOU

Of course, Sentibet is a demonstration of Neurolingo'a Mnemosyne technologies that may not seem to fit exactly with many enterprise applications. However, the types of problems solved by Neurolingo in the Sentibet project are a dramatic example of many of the types of solutions that Neurolingo has developed for the past seven years using its text analytics technologies at major banks, logistics companies and other major enterprises.

The key to Neurolingo's power is its ability to treat any source of unstructured text as a potential language unto itself – and to be able to identify languages even within those sources that require special treatment and processing. This allows you to apply Neurolingo technologies to not only very broad problems covering a variety of ways in which people express themselves in text but also very specific problems for very specific cohorts within a particular area of expertise or interest. There are many examples of how this could apply to major industries. Some relevant examples include:

- **Finance**
  - Trading triggers based on unstructured text data mining
  - Summarise and draw conclusions from stacks of reports
- **Sports**
  - Enrich social media offerings and offer improved odds-making input
  - Target team match-ups for media markets
- **Politics**
  - Interpret voter intents more accurately
  - Determine more rapidly when to adjust strategies
- **Health**
  - Interpret research and diagnostics for rapid response actions

Neurolingo can deploy its Mnemosyne platform in your own enterprise or maintain its text analytics capabilities as a cloud service, enabling you to focus on your business rather than on the ins and outs of maintaining semantic text analytics technologies. The choice is yours. Either way, you will be working with one of the most advanced sources of text analytics software available in the world today, enabling you to extract focus and intent from major sources of unstructured information with remarkable accuracy, ease and maintainability.

If you are looking for more information on how Neurolingo can help your enterprise to turn any source of content into a source of powerful insight, please content Neurolingo at:

www.neurolingo.com

ABOUT THE AUTHOR

John Blossom

President

Shore Communications Inc.

jblossom@shore.com

John Blossom is one of the most widely recognized content industry analysts, providing thought leadership to executives in search of new approaches to rapidly changing markets for publishing and technology products and services. Mr. Blossom founded Shore Communications Inc. in 1997, specializing in research and advisory services and strategic marketing consulting for publishers and content service providers in enterprise and media markets. Mr. Blossom's engagements have included strategic marketing consulting for major publishing and technology corporations and startups as well as speaking engagements at major conferences and advisory services for senior industry executives.

Mr. Blossom's career spans more than twenty years of marketing, research, product management and development in advanced information and media venues, including the marketing and development of financial information services at global financial publishers and financial services companies (Citicorp, Quotron and for Reuters Holdings PLC), as well as earlier experience in broadcast media. Mr. Blossom has served as a Director of Market Research for Risk/Waters Group and as Vice President and Lead Analyst at Outsell, Inc., where he provided research and analysis coverage of content technologies and financial and corporate information markets for major corporate clients, and developed successful online ecommerce services for research reports.

Mr. Blossom has been quoted in many major news and trade publications and media outlets, including The Wall Street Journal, CEO Magazine, BusinessNow, EContent Magazine, USA Today, the Financial Times, Information Today and Simba Research. Mr. Blossom's ContentBlogger weblog won the Software and Information Industry Association 2007 CODiE award for Best Media Blog. Mr. Blossom is the author of the book "Content Nation: Surviving and Thriving as Social Media Changes Our Work, Our Lives and Our Future," published by John Wiley & Sons, Inc. in January 2009. Mr. Blossom speaks regularly at industry conferences on topics related to the content industry and is a member of the Board of Directors of the Software and Information Industry Association Content Division.

Mr. Blossom's engagements include speaking and consulting in the U.S., U.K, Spain, Italy, India and Japan. His speaking engagements include appearances as a keynote speaker, panel moderator and panelist speaker at the SIIA Information Industry Summit,,SIIA NetGain, SIIA Financial Information Summit, SLA Annual Conference,The National Press Club, C-SPAN Television, The Commonwealth Club, ASIDIC, NFAIS, Buying and Selling eContent, Search Engine Strategies, Infovision (India), InfoCommerce Annual Conference, OCLC Symposium, TransPromo Annual Conference, Uchida Spectrum User Symposium (Tokyo).

Mr. Blossom's areas of expertise include media and enterprise content and technology thought leadership, social media, Web 2.0, knowledge management, online collaboration, market research, market analysis, business plan development and analysis, product analysis and development, e-books, mobile markets and user interface design and development.