Vendor Landscape: Big Data Text Analytics

Take Action From Customer Insights Hidden In Unstructured Data

by Boris Evelson

November 10, 2015 | Updated: November 11, 2015

Why Read This Report

Without the insights buried in petabytes of unstructured data, your customer view will never be complete. But most of the axioms, best practices, and technologies used for structured data do not apply to data mining and analysis of unstructured data. This report helps demystify text analytics processes for application development and delivery (AD&D) professionals. AD&D pros can also use the findings in this report to map their business requirements to vendor capabilities in the text analytics vendor selection process.

Key Takeaways

It's A Fragmented Vendor Landscape: Choose Based On Domain Expertise

Forrester uncovered a landscape with over 40 vendors but few differentiating features. While the underlying technology may be different from vendor to vendor, at the end of the day, most produce similar results. Choose your vendors based on industry vertical and business-domain-specific expertise, one of the last bastions of differentiation in text analytics.

Text Ingestion Is A Key Differentiator: Map Vendor Capabilities To Your Data Sources

Many of the vendors expect you to provide text for mining and analysis in a simple, cleansed text file. Choose vendors based on their capabilities to process complex data sources, such as long email chains with nested emails, attachments, and attachments within attachments such as compressed files.

Vendor Landscape: Big Data Text Analytics

Take Action From Customer Insights Hidden In Unstructured Data

by Boris Evelson
with Holger Kisker, Ph.D., Ian Jacobs, Maxie Schmidt-Subramanian, Sophia Christakis, and Ian McPherson

November 10, 2015 | Updated: November 11, 2015

Table Of Contents

Notes & Resources

Forrester interviewed 42 vendor and seven user companies.

Related Research Documents

Market Overview: Text Analytics

TechRadar™: Business Intelligence, Q1 2015

FORRESTER®

Forrester Research, Inc., 60 Acorn Park Drive, Cambridge, MA 02140 USA
+1 617-613-6000 | Fax: +1 617-613-5000 | forrester.com

Start Your Text Analytics Vendor Selection With Basic Technology

In our previous report on text analytics, we highlighted major use cases for customer-facing (such as voice of the customer [VoC] and market intelligence) and noncustomer-facing back office processes (such as document classification and risk management).[1] The market opportunity is huge, as most enterprises only use 27% of their semistructured and 31% of their unstructured data for business insights and decision-making.[2] For this vendor landscape report we screened more than150 text analytics providers in a diverse and highly fragmented market and selected the top 42 vendors for a detailed analysis. We categorized these text analytics vendors by whether they (see Figure 1).

- **Offer a platform, applications, or APIs.** In addition to the packaged and preconfigured text analytics applications, platforms provide buyers with the capabilities to extend the solution with custom developed applications fit for any business purpose. The vendor preconfigures and "trains" applications to serve a specific industry vertical or business domain (risk management, sentiment analysis, etc.) use cases.[3] APIs allow AD&D pros to further customize the platform and applications or embed text analytics into other transactional or analytical applications (see Figure 2).
- **Deliver a platform fully or partially built on open source technologies.** Open source software confers advantages and risks. Open source technologies are transparent (you can see the source code) and vendor independent (if one supplier drops support you can always find another). But some buyers, especially in highly regulated industries, or those building highly mission critical applications, have legitimate concerns that bugs or even malicious code can find its way into crowd-sourced source code. Many AD&D pros even deploy tools like Black Duck Software to identify and mitigate open source-related risks across their application portfolios.[4]
- **Provide a combination of products and services.** Most of the products reviewed in this research often require moderate to heavy involvement of vendor professional services organization to setup, customize, and train.
- **Sell to end users or sell OEM-embeddable components to ISVs.** Most of the vendors in this research boast a high degree of customization capabilities. One indirect way to judge the degree of customization offered by the platform is to look at the percentage of revenues the vendor derives from direct sales to the end users versus sales via indirect channels such as OEM-ing and embedding their platform into third-party applications. A high ratio of indirect/OEM sales is a good indication that the platform is highly customizable.
- **Are based on multitenant cloud or single tenant on-premises architecture.** Cloud-based text analytics platforms can help AD&D pros avoid the headaches often associated with deploying and maintaining on-premises software. Multitenant architecture (not applicable to platforms architected for single tenancy that just happen to be hosted in the cloud) is also very elastic, allowing on demand provisioning and deprovisioning of applications and computing resources. Forrester clients whose business and operational requirements call for total control of the application and release cycles, and who are concerned about the privacy and security issues, may choose an on premise solution.

FIGURE 1 Text Analytics Vendor Segmentation

							Revenues split by	
Vendor	**Product name**	**Open source compo-nents**	**Custo-mers**	**Text analytics revenue**	**Products/ services**	**Direct/ OEM**	**Cloud/on-premises deployment**	**Geographic regions NA/ Europe/ APAC/rest of the world**
Add-Structure	Scaffold	N	3	<$10M	95/5	100/0	100/0	50/0/50/0
Angoss Software	Knowledge Reader	N	100+	<$10M	75/25	100/0	0/100	100/0/0/0
Ascribe	Ascribe Intelligence	N	300	<$10M	90/10	90/10	100/0	50/30/10/10
Attensity	Attensity Analyze, Attensity Q, and Attensity Semantic Annotation SDK	N	50+	$10-$99M	80/20	100/0	20/80	80/20/0/0
Attivio	Attivio	N	45	$10-$99M	74/26	70/30	0/100	70/15/0/15
Basis Technology	Rosette SDK	N	52	$10-$99M	90/10	75/25	0/100	60/20/15/5
Bitext	Bitext Linguistic Analysis Engine	N	20+	<$10M	80/20	90/10	30/70	70/25/2/3
Brainspace	Brainspace Discovery, Brainspace for Enterprise	N	200+	$10-$99M	100/0	90/10	30/70	80/10/5/5
Cambridge Semantics	Anzo Unstruc-tured	N	75	<$10M	80/20	100/0	40/60	93/5/2/0
Clarabridge	Clara-bridge CX Suite	N	800+	$10-$99M	80/20	95/5	90/10	88/12/0/0
Content Analyst	CAAT 3.18	N	30+	$10-$99M	95/5	0/100	0/100	85/5/5/0

FIGURE 1 Text Analytics Vendor Segmentation (Cont.)

							Revenues split by	
Vendor	**Product name**	**Open source compo-nents**	**Custo-mers**	**Text analytics revenue**	**Products/ services**	**Direct/ OEM**	**Cloud/on-premises deployment**	**Geographic regions NA/ Europe/ APAC/rest of the world**
Revealed Context (Conver-seon)	ConveyAPI	Y	60	<$10M	70/30	50/50	100/0	90/10/0/0
Dell	Statistica	Y	375	<$10M	100/0	100/0	100/0	64/33/3/0
Digital Reasoning	Synthesys	N	20+	<$10M	80/20	100/0	0/100	70/30/0/0
EPAM	InfoNgen, Blueprint	Y	130+	$10-$99M	50/50	90/10	50/50	60/25/15/0
Etuma	Etuma Feedback Categorizer	N	25	<$10M	100/0	80/20	100/0	5/94/1/0
Expert System	Cogito Categorizer and Discover	N	100+	$10-$99M	40/60	95/5	10/90	25/75/0/0
FICO	FICO Text Analyzer	Y	10+	<$10M	80/20	100/0	0/100	100/0/0/0
Fractal Analytics	Dcrypt	Y	5	<$10M	80/20	100/0	0/100	80/20/0/0
Haystac	Indago	Y	5	<$10M	90/10	100/0	0/100	85/15/0/0
HPE	HP IDOL	N	100+	>$100M	50/50	100/0	0/100	60/30/5/5
IBM	IBM Social Media Analytics, IBM Watson Developer Cloud	N	100+	>$100M	50/50	100/0	100/0	50/30/10/10
Infegy	Infegy Linguistics	N	60	<$10M	100/0	100/0	80/20	85/10/5/0

FIGURE 1 Text Analytics Vendor Segmentation (Cont.)

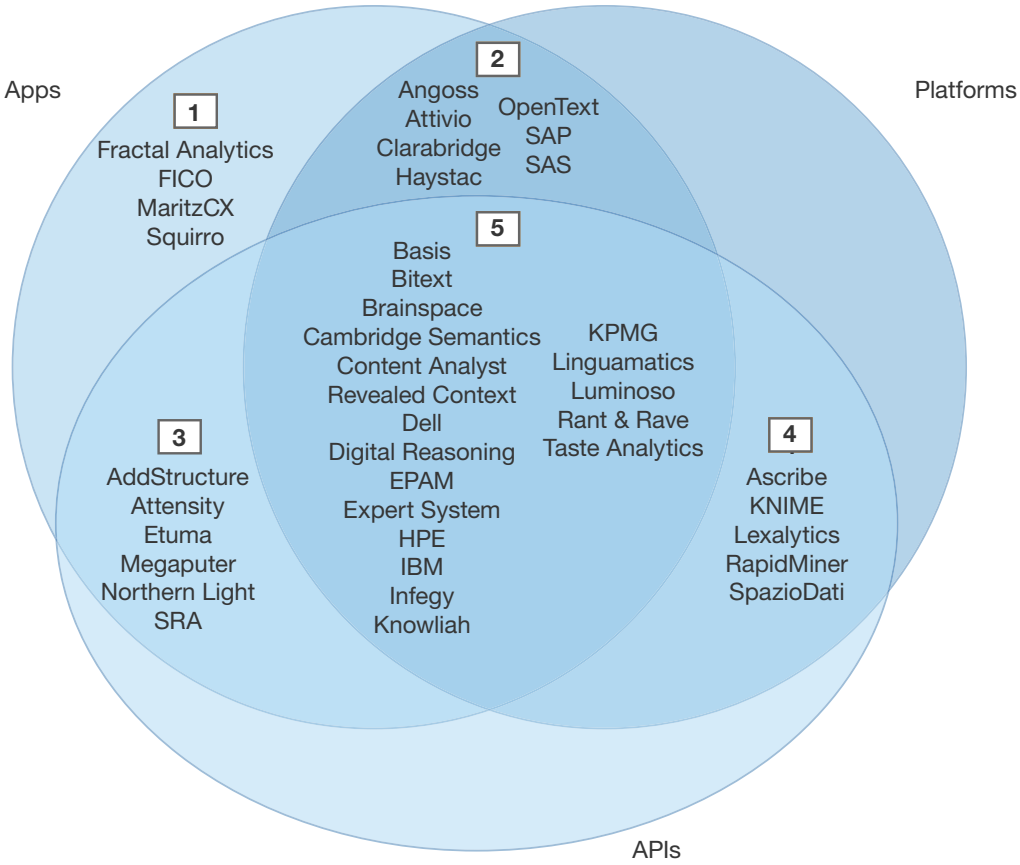
							Revenues split by	
Vendor	**Product name**	**Open source compo-nents**	**Custo-mers**	**Text analytics revenue**	**Products/ services**	**Direct/ OEM**	**Cloud/on-premises deployment**	**Geographic regions NA/ Europe/ APAC/rest of the world**
KNIME	KNIME Analytics Platform	Y	3000+	<$10M	80/20	50/50	0/100	40/40/18/2
Knowliah	Knowliah Enterprise Search & Analytics	N	24	<$10M	70/30	90/10	20/80	10/85/5/0
KPMG	KPMG Advanced Text Analytics	N	25+	$10-$99M	0/100	100/0	20/80	70/15/10/5
Lexalytics	Salience, Semantria, Semantria Excel Add-In	N	150	<$10M	90/10	20/80	25/75	80/10/5/5
Linguama-tics	I2E	N	50+	$10-$99M	80/20	100/0	50/50	70/25/5/0
Luminoso	Luminoso Analytics, Luminoso Compass, Luminoso API	Y	130	<$10M	80/20	100/0	98/2	90/5/5/0
MaritzCX	MaritzCX Text Analytics	N	65+	<$10M	95/5	100/0	100/0	90/10/0/0
Meaning Cloud	Meaning Cloud	Y	50+	<$10M	40/60	90/10	50/50	40/30/10/20
Megaputer Intelligence	PolyAnalyst 6.5	N	70	<$10M	72/28	100/0	0/100	77/16/5/2
Northern Light	MI Analyst	N	35	<$10M	0/100	98/2	100/0	75/25/0/0

FIGURE 1 Text Analytics Vendor Segmentation (Cont.)

							Revenues split by	
Vendor	**Product name**	**Open source compo-nents**	**Custo-mers**	**Text analytics revenue**	**Products/ services**	**Direct/ OEM**	**Cloud/on-premises deployment**	**Geographic regions NA/ Europe/ APAC/rest of the world**
OpenText	OpenText Discovery Suite	Y	100+	$10-$99M	50/50	90/10	0/100	45/40/15/0
Rant & Rave	Rant & Rave Platform	N	40	$10-$99M	0/100	100/0	100/0	2/98/0/0
RapidMiner	RapidMiner Studio, RapidMiner Server, RapidMiner Text Processing	Y	330	<$10M	80/20	90/10	25/75	45/40/12/3
SAP	SAP Hana	N	6000+	$10-$99M	90/10	100/0	0/100	60/30/5/5
SAS	SAS Contextual Analysis & SAS Text Miner	N	1000+	$10-$99M	90/10	99/1	0/100	45/32/18/5
SpazioDati	Dandelion API	Y	15	<$10M	99/1	90/10	90/10	10/90/0/0
Squirro	Squirro	Y	24	<$10M	60/40	100/0	50/50	40/50/5/5
SRA International	NetOwl	N	50+	<$10M	100/0	100/0	20/80	70/20/0/10
Taste Analytics	Stratifyd Signals	N	28	<$10M	100/0	100/0	100/0	80/20/0/0

Some entries are Forrester estimates

FIGURE 2 Text Analytics Vendor Segmentation By Platform, Applications, And APIs



Vendor segment	Description	Platform	Application	APIs
1. Applications only	Industry vertical and business-domain-specific applications		X	
2. Platforms/ applications	General purpose platforms that can be used for multiple industry vertical and business domain specific applications	X	X	
3. Customizable applications	Applications that can be customized and extended with APIs		X	X
4. Customizable platforms	General purpose platforms that can be further customized and extended with APIs	X		X
5. General purpose customizable platforms	General purpose platforms that can be used for multiple industry vertical and business-domain-specific applications and further customized and extended with APIs	X	X	X

Shortlist Text Analytics Vendors Based On Business Domain Expertise

While text analytics, just like any other automation or programmatic technology, attempts to eliminate dependency on human resources, the market is not quite there yet. All text analytics applications require some level of domain-specific training and therefore necessitate involvement of subject matter human resources. Forrester recommends that AD&D pros working on text analytics vendor selection further shortlist vendors based on their vertical industry and business domain expertise (see Figure 3 and see Figure 4).[5] Look for such expertise based on the availability of prepackaged applications, including domain-specific ontologies.

FIGURE 3 Text Analytics Vendors Industry Vertical Specialization

Vendor name	Manufacturing	Retail and wholesale	Business services and construction	Media, entertainment, and leisure	Utilities and telecommunications	Financial services and insurance	Public sector and healthcare
AddStructure		X	X	X			
Angoss Software		X		X	X	X	
Ascribe	X	X	X	X	X	X	X
Attensity	X	X		X	X	X	
Attivio	X	X		X	X	X	X
Basis Technology	X	X	X	X		X	X
Bitext	X	X		X	X		
Brainspace	X		X			X	X
Cambridge Semantics	X	X	X	X	X	X	X
Clarabridge	X	X		X	X	X	X
Content Analyst	X		X				X
Revealed Context (Converseon)	X	X		X	X	X	X
Dell	X	X	X	X	X	X	X
Digital Reasoning						X	X
EPAM	X	X	X	X	X	X	X
Etuma	X	X	X	X	X	X	X
Expert System	X	X		X	X	X	X
FICO	X	X	X	X	X	X	X
Fractal Analytics	X	X				X	X
Haystac	X		X	X	X	X	X

FIGURE 3 Text Analytics Vendors Industry Vertical Specialization (Cont.)

Vendor name	Manufacturing	Retail and wholesale	Business services and construction	Media, entertainment, and leisure	Utilities and telecommunications	Financial services and insurance	Public sector and healthcare
HPE	X	X	X	X	X	X	X
IBM	X	X	X	X	X	X	X
Infegy	X	X	X	X	X	X	X
KNIME	X	X	X	X	X	X	X
Knowliah	X	X	X	X	X	X	X
KPMG	X	X	X	X	X	X	X
Lexalytics	X	X	X	X	X	X	X
Linguamatics	X						X
Luminoso	X	X	X	X	X	X	X
MaritzCX	X	X	X		X	X	X
MeaningCloud	X	X		X	X	X	X
Megaputer Intelligence	X	X	X			X	X
Northern Light	X	X	X		X	X	X
OpenText	X	X		X		X	X
Rant & Rave	X	X	X	X	X	X	X
RapidMiner	X	X	X	X	X	X	X
SAP	X	X	X	X	X	X	X
SAS	X	X	X	X	X	X	X
SpazioDati	X	X	X	X	X	X	X
Squirro	X				X	X	
SRA International				X		X	X
Taste Analytics	X	X		X	X	X	X

Some entries are Forrester estimates

FIGURE 4 Text Analytics Vendors Business Domain Specialization

Vendor name	Document classification	Legal	Competitive/ market intelligence	Risk management/ fraud detection	Semantic search	Sentiment analysis	Social media monitoring	Voice of the customer	Voice of the employee
AddStructure	X				X	X		X	X
Angoss Software			X	X		X		X	
Ascribe			X			X		X	X
Attensity			X	X		X	X	X	
Attivio	X		X	X		X	X	X	X
Basis Technology	X	X	X	X	X	X	X	X	X
Bitext	X					X	X	X	X
Brainspace	X	X		X	X	X			
Cambridge Semantics	X		X	X	X	X	X	X	X
Clarabridge			X	X	X	X	X	X	X
Content Analyst		X	X		X		X	X	
Revealed Context (Converseon)			X			X	X	X	X
Dell			X	X	X	X	X	X	X
Digital Reasoning			X	X	X	X			
EPAM	X	X	X		X	X	X	X	
Etuma			X		X	X	X	X	X
Expert System	X	X	X	X	X	X	X	X	X
FICO				X	X	X		X	
Fractal Analytics	X	X	X	X	X	X	X		
Haystac	X	X			X				

Some entries are Forrester estimates

FIGURE 4 Text Analytics Vendors Business Domain Specialization (Cont.)

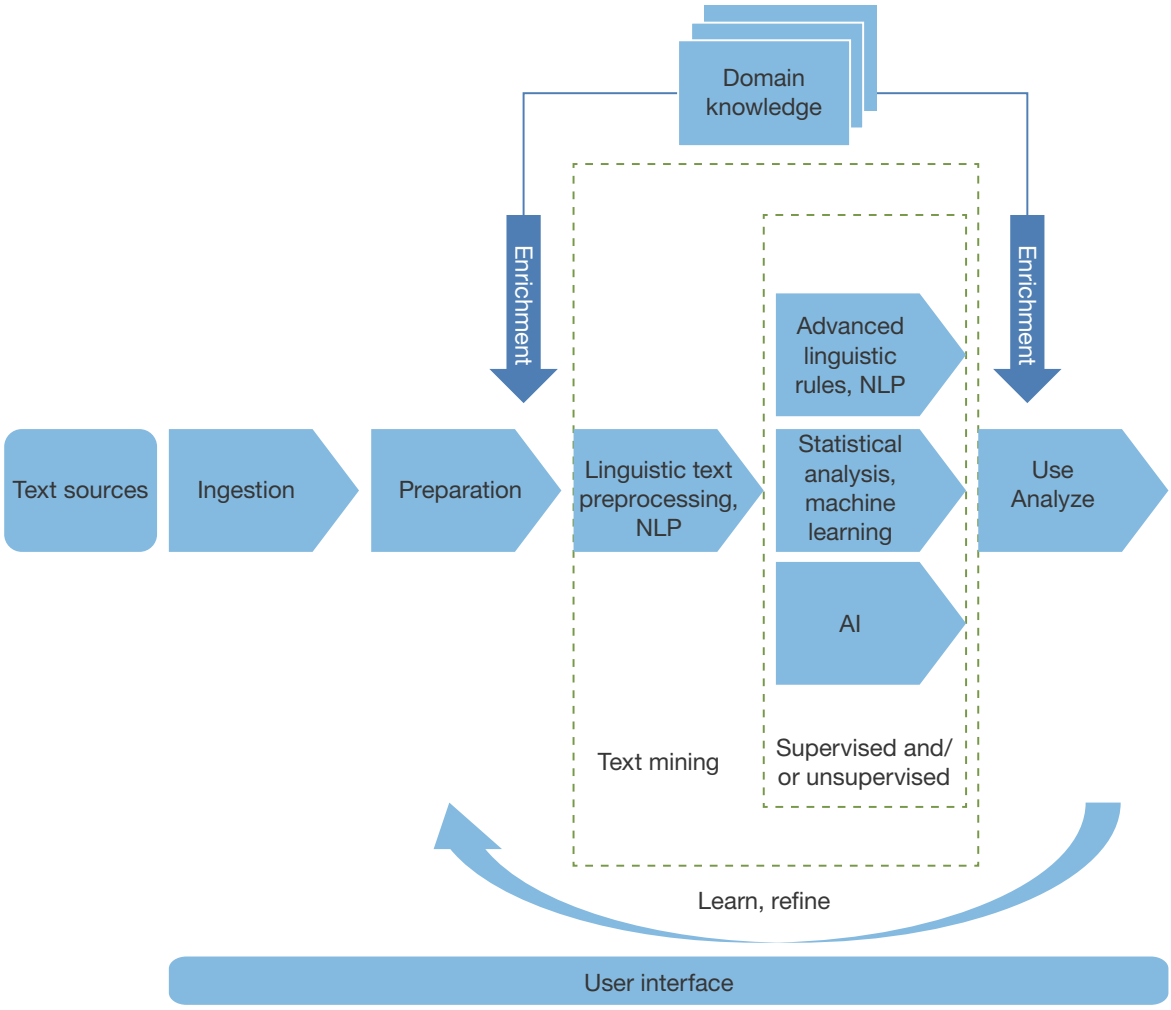
Vendor name	Document classification	Legal	Competitive/ market intelligence	Risk management/ fraud detection	Semantic search	Sentiment analysis	Social media monitoring	Voice of the customer	Voice of the employee
HPE	X	X	X	X	X	X	X	X	X
IBM	X	X	X	X	X	X	X	X	X
Infegy	X		X	X		X	X	X	
KNIME	X	X	X	X	X	X	X	X	X
Knowliah	X	X	X	X	X	X	X	X	X
KPMG	X	X	X	X	X	X	X	X	
Lexalytics	X	X	X	X	X	X	X	X	X
Linguamatics	X		X	X	X	X	X	X	
Luminoso	X	X	X	X	X	X	X	X	X
MaritzCX	X		X			X	X	X	X
MeaningCloud	X		X		X	X	X	X	X
Megaputer Intelligence	X	X	X	X		X	X	X	X
Northern Light			X		X	X	X		
OpenText	X		X		X	X			
Rant & Rave	X		X	X	X	X	X	X	X
RapidMiner	X	X		X	X	X	X	X	X
SAP	X	X	X	X	X	X	X	X	X
SAS	X	X	X	X	X	X	X	X	X
SpazioDati	X		X	X	X	X	X	X	
Squirro			X	X					
SRA International	X		X	X	X	X	X	X	
Taste Analytics	X		X	X	X	X	X	X	X

Some entries are Forrester estimates

Deconstruct The Text Analytics Black Box To Reveal Vendor Differences

To most business users, text analytics is a black box where unstructured text goes in and keywords, sentiments, and other structured information come out. AD&D pros have to make more informed platform buying decisions and therefore need to deconstruct the black box to look deeper into the specific process steps and capabilities. Understanding the process workflow, the components in the workflow, and the specific functionality of each component is a key to mapping vendor capabilities to each specific use case (see Figure 5).

FIGURE 5 Text Analytics Components And Process Flow



Extract, Ingest, Digitize, And Prepare The Text For Mining

The process starts with connecting applications to data sources, ingesting text, and preparing the text for the next step — the actual analysis. Vendor capabilities vary significantly: some provide advanced out-of-the box features; others rely on third-party data integration tools with text processing functionality; still others require professional services to customize the data ingestion components. AD&D pros should evaluate the text ingestion features of text analytics platforms such as (see Figure 6):

› **Connectivity to a variety of data sources.** Unstructured text exists in variety of sources and formats, each of which presents its own challenges. Typical text sources include documents stored in file shares, archives, and ECM repositories; eBooks, emails, web pages, social streams and chat

applications; as well as freeform text entry into desktop, enterprise resource planning, customer relationship management, and analytics applications. Many sources require special treatment before the text can be processed; for example, emails present a very specific challenge due to the potential for multiple levels of email chains and nested attachments (especially when attachments are compressed files). Another key platform differentiation is a capability to "understand" and process complex document forms such as invoices and purchase orders.

- **Text ingestion and conversion.** AD&D pros need to consider whether they need to perform text analytics on a steady data source with a low frequency of changes and updates, such as a library of documents like patent filings, or data sources that change frequently, such as online news stories. In the former case, reanalyzing the entire data set after each new update is not efficient, so AD&D pros should look for the ability to handle incremental or "delta" updates and process updates to data sources like social media streams or chats in real time. Delta updates can also provide additional metadata for temporally tagging documents (tracking what content changed over time). Furthermore, not all data sources provide text in a digitized format, so one must consider text analytics software capabilities to convert audio (speech-to-text) or hard-copy (OCR) data to digital text format.[6]
- **Text preprocessing and preparation.** After ingesting and digitizing text, the process formats the text for mining. This step aims to generate a machine-processable structure such as CSV, XML, JSON, or a database records. Some of the functions performed during this step include filtering out unneeded text such as headers, footers, and legal disclaimers and performing basic calculations such as word counts. Some of the latest technologies attempt to bypass such "preparatory" steps and analyze text data as is.
- **Text cleansing.** A subset of the text preparation step is a capability to clean the text. For example, Infegy and a few other text analytics platforms can automatically find and delete embedded JavaScripts (often found embedded in web pages) and fix malformed HTML (missing tags).
- **In-place text mining.** Rather than moving the documents to its platform (replicating storage and potentially exposing highly sensitive documents to an outside platform), Haystac can take a digital snapshot of the documents, which is all it needs to classify, categorize, cluster, and summarize documents on its own platform without physically moving the actual documents in and out of the client repositories.

FIGURE 6 Text Ingestion Components

Text ingestion components	Examples of text analytics platforms features
Text source connectors	Database via ODBC/JDBC, crawlers (DAM, ECM, file system, IMAP, news, RSS, social media, websites), cloud storage, FTP
Input formats	Documents, complex forms (invoices, etc.), eBooks, emails, attachments, compressed files, surveys, SAS files, SPSS files, R files, JSON, OWL, RDF, XML
Text ingestion	Batch, streaming, voice to text, OCR, capability to handle delta updates
Cleaning	Formatting, headers, footers, annotations, embedded Javascript, malformed HTML

Map Your Use Cases To Linguistic, Statistical, Trained, And Unsupervised Techniques

The actual mining of the text comes next. Text mining — often used synonymously with text analytics — uses linguistic and statistical rules, algorithms, and techniques, frequently working together in a complementary fashion, to find structures and patterns in text. The components that specifically deal with understanding and processing human languages are categorized as natural language processing (NLP). Text analytics process moves on with the steps to (see Figure 7):

- **Autodetect and analyze text in multiple languages.** Text analytics is not created equal for every language. While most of the vendors identified in this research can automatically detect and process text in multiple languages, their capabilities differ widely. Linguistic rules are language-dependent. AD&D pros should ask the vendors to disclose not only which languages they support, but also whether all linguistic rules are available for each supported language. Investigate how the vendor handles language changes such as slang or foreign words that become part of another language. Vendors often place their multilanguage capabilities into tier-one (support for all linguistic rules and statistical analysis) and tier-two (support only a few linguistic rules or only statistical analysis) languages. Ask about text analytics platform features such as processing documents in multiple languages, documents that contain different languages, and advanced autotranslation capabilities.
- **Preprocess text using linguistic rules and NLP.** This step prepares text for more advanced analysis. There are rules for identifying parts of speech; identifying synonyms, roots, and word stems; identifying roles (subject, object, action, etc.); and identifying lexical chains (word sequences) and word nets (associating words to help determine their context). One of the differentiating features of this step is tokenization or end of sentence detection (which is as not as simple as just looking for periods). Another differentiating feature is a capability to automatically fix incorrect grammar.
- **Analyze text using advanced linguistics rules and statistical algorithms.** One of the advantages of statistical language analysis is that it's mostly language-independent, based on parameters like word proximities and sequences. Advanced linguistic rules-based analysis (including NLP) and/or statistical analysis can perform sub-document, document, and cross document type

analysis (see Figure 8). Sub-document level text analytics capabilities may include extracting entities and concepts, detecting events, uncovering lexical chains and word sequences, mapping entity relationships, extracting sentiment, and tagging (temporal and spatial). Document-level analysis may include document categorization and classification and template matching as well as automatically producing document summaries. Cross-document text analytics usually provides document clustering capabilities, duplicate and near duplicate document detection, and cross-document entity and concept resolution.

› **Analyze text using supervised and unsupervised techniques.** Both linguistic and statistical language analysis may use supervised ("trained") or unsupervised approaches. Supervised analysis starts with a base sample of documents fed into the system; subsequent analysis ascertains the degree of similarity or the difference between the base and all other documents. This works best in situations such as investigations or eDiscovery when subject-matter experts (SMEs) code or tag sample document sets (often referred to as a "golden copy"). In contrast, unsupervised analysis runs loose against any collection or stream of text without first understanding its structure or content. While an unsupervised approach requires less preparation and maintenance, it can potentially be less accurate. To address the strengths and weaknesses of both approaches most modern text analytics platforms offer a combination of supervised and unsupervised techniques.

  When mapping a specific text analytics platform to your use case, remember that supervised techniques will not just miraculously produce results. Rather, they require human resources with relevant subject matter expertise. AD&D pros should also ask text analytics vendors whether the rules built after training the application can be exported, imported, and reused in other applications without the need to retrain the model.

› **Go beyond sentiment with emotion and intention analysis.** Echoing Forrester research on turning insights into action, some text analytics vendors are going beyond classification, sentiment, and entity extraction with "actionability analysis."[7] Identifying customer sentiment (positive, negative, neutral) is a step in the right direction to start addressing customer concerns. Overlay sentiment with emotion analysis (happy, sad, confused, angry) and a CI pro is now armed with a more precise insight as to how to change customer sentiment by addressing their emotional state. Add in intention analysis (intent to buy, intent to switch to another brand) to the mix and a CI pro now has razor sharp insight on whether to adjust the campaign message, fix product issues, or make a preemptive offer to the customer.

› **Dig deeper into auto classification of text with artificial intelligence.** Also known as cognitive computing or deep learning, the latest advances in neural network-based AI allow text analytics platforms to uncover more complex patterns such as analogical relationships (e.g., short is to shortest as big is to biggest) and linguistic regularities (e.g., "king" + "woman" is close to "queen").[8] Cognitive computing can also attempt to automatically build word classes and taxonomies (e.g., "carnivore" and "cormorant" both relate to animals). Beware: AI is still in its infancy and is by no means plug-and-play. Real battle-hardened, mission-critical AI applications still require significant investments in training these systems with domain knowledge and expertise.

FIGURE 7 Text Analytics Components And Capabilities

Text analytics components and capabilities	Examples of text analytics platform features
Multilanguage capabilities	Automatic language detection, document sets containing documents in different languages, multiple languages in a document, autotranslation
Linguistic text preprocessing, NLP	Parsing, syntactic parsing, tokenization, end of sentence detection, part of speech identification, word stemming, sequence of stems (lemmas), roles (object subject, action, etc.), spelling correction, synonyms
Sub-document-level text analytics	Concept extraction, co-reference resolution (linking multiple mentions of the same entity, concept), entity extraction, event detection, lexical chains, word sequences, relationship mapping, sentiment extraction, word nets (associating words to help in disambiguation to determine the contextual sense of the terms), emotion detection, spam detection, intensity detection, intention detection
Document-level text analytics	Categorization/classification, including multilevel hierarchical folders, summarization, template matching
Cross-document text analytics	Clusters, near duplicate detection, cross-document entity and concept resolution, document similarity
Tagging	Temporal, spatial, entity
Relationship maps	Documents, entities, topics, concepts, keywords
Advanced functionality	Neural networks, forecasts, logistic regression, decision trees, analogical relationships, linguistic regularities, automatically building word classes and taxonomies

FIGURE 8 Linguistic Versus Statistical Analysis; Supervised Versus Unsupervised Text Analytics Vendor Platforms

Vendor name	Supervised and/or unsupervised techniques	Linguistic rules and/or statistics based engine
Angoss Software	Both	Both
Ascribe	Both	Both
Attensity	Both	Both
Attivio	Both	Both
Basis Technology	Both	Both
Bitext	Both	Linguistic rules
Brainspace	Both	Both
Cambridge Semantics	Both	Both
Clarabridge	Both	Both
Content Analyst	Both	Statistical processing
Revealed Context (Converseon)	Both	Both
Dell	Both	Both
Digital Reasoning	Both	Both
EPAM	Both	Both
Etuma	Both	Both
Expert System	Both	Both
FICO	Both	Both
Fractal Analytics	Both	Both
Haystac	Both	Statistical processing
HPE	Both	Both

FIGURE 8 Linguistic Versus Statistical Analysis; Supervised Versus Unsupervised Text Analytics Vendor Platforms (Cont.)

Vendor name	Supervised and/or unsupervised techniques	Linguistic rules and/or statistics based engine
IBM	Both	Both
Infegy	Both	Both
KNIME	Both	Both
Knowliah	Supervised	Both
KPMG	Both	Both
Lexalytics	Both	Both
Linguamatics	Supervised	Both
Luminoso	Both	Both
MaritzCX	Both	Both
MeaningCloud	Both	Both
Megaputer Intelligence	Both	Both
Northern Light	Supervised	Linguistic rules
OpenText	Both	Both
Rant & Rave	Both	Both
RapidMiner	Both	Both
SAP	Both	Both
SAS	Both	Both
SpazioDati	Both	Both
Squirro	Both	Statistical processing
SRA International	Both	Both
Taste Analytics	Unsupervised	Both

Some entries are Forrester estimates

Enrich, Interact With, And Verify Accuracy Of Your Text Analytics Applications

Text analytics is never a single step process. It requires multiple trial and error iterations and therefore necessitates a rich user interface (UI) to conduct the experiments and continually improve the system. AD&D pros should consider additional text analytics components to include (see Figure 9):

› **UI for interacting with the data.** Other than text analytics tools that only provide APIs, most text analytics platforms and applications require some kind of a presentation layer — a UI — to search, browse (through hierarchies, for example), analyze, refine results or use case parameters, and otherwise interact with the text analytics process and data. A use case may even call for browsing through unprocessed text to peek into patterns, anomalies, outliers, or any other indicators that

may give a user ideas on what to look for in the text and what kind of text-mining techniques to use. Searching and browsing functionality is often based on query languages such as KDTL ("knowledge discovery in text language"), XQuery for semistructured text represented in XML format, or SPARQL, a semantic query language for RDF triple store databases.

- **Post-processing and data enrichment with domain knowledge.** Once the keywords, entities, concepts, roles, and other semantic tags have been identified and extracted, they may be refined and further classified. For example, a use case may call for enriching the findings with business domain- and industry-specific ontologies (bodies of knowledge that define topics and concepts for a particular domain), taxonomies (relationships like hierarchies of elements within a domain), and lexicons (dictionaries of all relevant terms within an ontology).[9] The enriched data may then be sent back for another cycle of text mining to derive refined or additional information and patterns. Such extra metadata tagging is a key enabler of semantic search — improving search accuracy by understanding the searcher's intent and the contextual meaning of terms — and the Semantic Web, both of which are widely used in knowledge management applications.[10]
- **Data enrichment with knowledge graphs.** In addition to domain specific ontologies, a few text analytics vendors now offer data enrichment and refinement via knowledge graphs. The knowledge graph was originally introduced by Google as knowledge base used to enhance its search with domain-specific knowledge of people, places, events, etc.[11] Basically, knowledge graphs are databases of "all things." Rather than constantly jumping to multiple websites checking information to refine or enhance results of a text mining process, knowledge graphs embed such information right into the results of the text mining process output.
- **Automated accuracy verification.** Most text analytics vendors claim 80% to 95% accuracy. Unfortunately, most of them verify the accuracy manually (by comparing manual tagging of text by linguistic scientists to the results from the automated text mining process) or by various statistical cross-validation techniques (K-fold, N-fold) that unfortunately only apply to statistical, not linguistic text analysis. Text analytics use cases that call for classification of a large number of documents can compare text analytics results against standard document templates, but such approach is not applicable to streaming text analytics that analyzes news feeds or social media. Infegy offers an original approach: compare text sentiment derived from product review comments on eCommerce sites like Amazon to the number of stars the reviewer gave that product. This approach effectively crowdsources accuracy verification.

FIGURE 9 Other Capabilities Of Text Analytics Platforms

Text analytics components and capabilities	Examples of text analytics platform features
Data enrichment	Ontologies (including taxonomies and lexicons), knowledge graphs, OWL support
Accuracy verification	Manual, K-fold and N-fold cross validation, F1 scores, comparison against standard corpus
UI options	GUI, scripts, command line
UI operations	Browsing, searching, query construction and refinement, input conditions, concepts and filters, create, explore, manipulate taxonomies
Text query languages	Xquery, KTDL, SPARQL

Properly Deployed, Text Analytics Can Have Tremendous Benefits

Deploying efficient and effective text analytics applications is not a simple walk in the park. But those who can overcome the challenges achieve tremendous tangible benefits, such as:

- **Improving customer experience.** Orbitz wanted to develop an enterprisewide customer focus to start making decisions based on real evidence. Using a text mining platform from Clarabridge, Orbitz was able to analyze multiple VoC data streams to determine and prioritize customer wishes. The result: a new customer loyalty program that earned Orbitz the ranking of No. 1 travel website in overall customer satisfaction by the American Customer Satisfaction Index.
- **Cost savings and efficiency gains.** A leading European telecom provider was overwhelmed by customer requests (800,000 emails per month) and needed to improve the efficiency by responding faster and more accurately through their support desk. Using text analytics technology from RapidMiner, the provider achieved a 70% reduction in categorization work and reduced the error rate for more than 50 different categories to less than 5%, achieving an ROI in a few months. In another example, a Big 4 consulting firm was struggling to analyze and extract insight from over 8000 long-form, natural language surveys. This process was historically brute force, requiring 75 consultants and 30 days to read, review, and report on findings. Using Brainspace Discovery with two hours of training, five consultants were able to complete the project in five days, producing superior results and reducing information leakage, saving over $880,000.
- **Avoiding M&A risk.** An acquisition team at a multibillion-dollar global pharmaceutical manufacturer used Northern Light SinglePoint strategic research portal — this particular company's repository integrated 15 market research sources comprising some 50 million documents. The team was able to find certain documents that challenged the market forecast for the technology of the acquisition target. As a result, the deal was cancelled and the company saved $100 million on what could have been a disastrous acquisition.

- **Early warning systems detecting risks and threats to human health.** At the Global Public Health Intelligence Network (GPHIN), 80 multilingual analysts used to manually comb through content from multiple sources worldwide (on average 4000 articles per day, but with peaks at 20,000-plus when a potential threat is detected). OpenText's text analytics capabilities partially automate the task by mining articles for a thousand or so concepts such as "mysterious ailments" and "outbreak." GPHIN system then assigns relevancy scores so analysts can review only the relevant pieces of information. As a result the system helps analysts count and track instances of possible threats and triggers responses — such as World Health Organization (WHO) declaring H1N1 a geographic pandemic, which in turn hurries the development of vaccines.
- **Resolving quality issues.** A leading PC manufacturer needed to discover quality issues more quickly and understand customer desires more thoroughly. With SAS Text Analytics the manufacturer was able to detect product issues in half the time of traditional warranty analysis and reduce warranty costs by 10% to 15%. With proactive changes to customer services and provided documentation based on the insights from the analysis of customer complaints, the manufacturer also achieved a 30 percent reduction in general information calls to the contact center.
- **Improving customer retention and opportunities for cross-sell and upsell.** A financial company had an inadequate process for evaluating its associates' messages to customers for consistency and compliance with professional standards. The company was able to evaluate only 2.5% of over 1.2 million messages every year. Using a new system based on the Megaputer Intelligence platform the new application extracted features of interest from all associates' messages and scored each message against core competencies such as empathy, professionalism, and correctness of response. The company estimates that the new application creates an estimated annual value of $11.8 million for the company a 7% increase in customers' willingness to recommend the company.

Recommendations

Simplify Text Analytics Vendor Selection With A Pragmatic Approach

The text analytics market has produced a diverse, fragmented, and overlapping vendor landscape. AD&D pros choosing a text analytics platform or an application should approach the selection process pragmatically and:

- **First consider the current enterprise software providers.** IBM, Hewlett Packard Enterprise, OpenText, SAP, and SAS offer broad and comprehensive text analytics platforms. If your relationship with these providers in other areas of enterprise software, applications, and services is solid put them at the top of your list.
- **Go with professional services providers if you know you'll need help.** A combination of products and services from a single provider works great when your own organization lacks subject matter experts and you only want a single finger to point at someone when a project goes south. If that's the case, consider EPAM, IBM, Hewlett Packard Enterprise, and KPMG.

› **Give preference to well-established vendors if your project is broad and global.** Don't be shy about engaging with small startups with a few clients and narrow regional presence when their technology works best for your specific use case. But for broad text analytics use cases, multilingual capabilities, and global support, leverage vendor profiles from this research to make sure you engage with larger vendors with a critical mass of customers and global presence.

Supplemental Material

Companies Interviewed For This Report

AddStructure
Angoss Software
Ascribe
Attensity
Attivio
Basis Technology
Bitext
Brainspace
Cambridge Semantics
Clarabridge
Content Analyst
Dell
Digital Reasoning
EPAM

Etuma
Expert System
FICO
Fractal Analytics
Haystac
Hewlett Packard Enterprise
IBM
Infegy
KNIME
Knowliah
KPMG
Lexalytics
Linguamatics
Luminoso
MaritzCX
MeaningCloud
Megaputer Intelligence
Northern Light
OpenText
Rant & Rave
RapidMiner
Revealed Context (Converseon)
SAP
SAS
SpazioDati
Squirro
SRA International
Taste Analytics

Endnotes

[1] Most organizations still analyze structured and unstructured data in silos, using different tools and serving different use cases — even as techniques such as statistical analysis, machine learning, natural language processing, and artificial intelligence are now bringing text analytics closer to the familiar world of BI. This report helps application development and delivery (AD&D) pros working on BI initiatives that include unstructured data sources to demystify text analytics by describing its typical components and process flows. See the "Market Overview: Text Analytics" Forrester report.

[2] These are estimates calculated using midpoints of ranges and are not exact. Source: Forrester's Global Business Technographics® Data And Analytics Survey, 2015.

[3] Supervised analysis starts with a base sample of documents fed into the system; subsequent analysis ascertains the degree of similarity or the difference between the base and all other documents.

[4] Software composition analysis is an emerging market, a cross section of the most important technologies from open source scanning vendors and traditional security assessment offerings. Code license management and software vulnerability discovery are coming together to provide value to enterprises attempting to secure and manage their software supply chain. See the "Vendor Landscape: Software Composition Analysis" Forrester report.

[5] We define the text analytics business domain specializations in Figure 4 as follows:

Document classification: The process of searching, finding, identifying, and classifying large volumes of data. See the "Market Overview: Text Analytics" Forrester report.

Legal: Refers to mining electronically stored information (ESI) related to lawsuits to identify broken processes, products, or services. The identification of case relevant, privileged, personal, and confidential documentation, or the

ability to categorize and metatag documents for storage in enterprise content management (ECM) repositories and assess them for relevance during eDiscovery. See the "Thirteen Legal Hold Tools And How To Use Them" Forrester report, see the "Understand The Context Of eDiscovery Tools For Your Enterprise" Forrester report, and see the "Market Overview: Text Analytics" report.

Competitive/market intelligence: The ability to obtain a deep understanding of competitor's assets, talent, capabilities, structures, processes, and culture, as well as identifying potential competitors, disruptive business models, products, and strategies before they gain momentum. See the "Proactive Competitive Market Intelligence Builds Competitive Advantage" Forrester report.

Risk management/fraud detection: Mining information from typically unused data sources (unstructured and triplet data) to identify potentially fraudulent patterns and trends. See the "Big Data In Fraud Management: Variety Leads To Value And Improved Customer Experience" Forrester report.

Semantic search: Standardized approach for discovering, querying, browsing, analyzing, visualizing, and modeling semantic constructs. Using text analytics to improve search functionality based on previous searches, content, topics, etc. See the "Meanings Matter: The Semantic Web Enriches Data Management And Fuels Processes" Forrester report.

Sentiment analysis: Measuring how many customer mentions of a topic were negative, positive, or neutral. A sentiment analysis tool that can identify sentiment on a topic or comment level, as opposed to just on a post level. See the "Q&A: The Social Analyst's Primer On VoC Programs, Data, Vendors, And Collaboration Tips" Forrester report.

Social media monitoring: Tools to help public relations professionals keep tabs on crises to enterprise listening platforms that provide monitoring plus analysis on social data to calculate sentiment, influencer rankings, and campaign effectiveness. See the "The Forrester Wave™: Enterprise Listening Platforms, Q1 2014" Forrester report.

Voice of the customer: Mining customer feedback from surveys, social media, emails, and calls and analyzing that feedback for insights to be incorporated into business decisions. See the "Voice Of The Customer Vendor Landscape, 2014" Forrester report.

Voice of the employee: Mining feedback from employees or partners that pertains to their ability to deliver great customer experiences and analyzing that feedback for insights to be incorporated into business decisions. See the "Cure Broken Customer Experiences With Voice Of The Employee Programs" Forrester report.

[6] Vendors of real-time speech analytics tools promise to allow companies to intervene at the moment of truth, while the customer and the contact center agent are still talking. This brief discusses the hurdles application development and delivery (AD&D) pros will need to overcome to justify the expenditure on this technology and the steps they will need to take to prepare for a world of alerts generated in real-time based on customer conversations. See the "Brief: Real-Time Speech Analytics — Still More Sizzle Than Steak" Forrester report.

[7] This report connects the dots between our research on BI, Agile BI, and big data; proposes best practices for merging previously separate efforts into a more cohesive systems of insight strategy; and offers actionable advice for AD&D pros working on BI and big data initiatives on supercharging BI and upgrading to 21st-century systems of insight. See the "It's Time To Upgrade Business Intelligence To Systems Of Insight" Forrester report.

[8] Forrester defines artificial intelligence (AI) as the theory and capabilities that strive to mimic human intelligence through experience and learning. AI capabilities include elaborate reasoning models to answer intricate questions and solve complex problems. Enterprise developers are starting to use AI to build cognitive computing systems. This report will help enterprise architecture professionals navigate the AI market landscape, understand how cognitive computing can enhance their business applications, and map how AI technologies may fit (or be retrofitted) in their existing architecture. See the "Artificial Intelligence Can Finally Unleash Your Business Applications' Creativity" Forrester report.

[9] Ontologies are bodies of knowledge that define topics and concepts for a particular domain and taxonomies are relationships like hierarchies of elements within a domain.

[10] The Semantic Web is a distributed collection of "linked data" on the Web. Just as web pages are linked together via hypertext, the goal of the Semantic Web is to link all the available public data with the purpose of making searches more effective. Source: "Encyclopedia: Definition of: Semantic Web," PCMag Digital (http://www.pcmag.com/encyclopedia/term/51092/semantic-web).

[11] The knowledge graph enables you to search for things, people or places that Google knows about and instantly get information that's relevant to your query. This is a critical first step towards building the next generation of search, which taps into the collective intelligence of the Web and understands the world a more like people do. Source: Amit Singhal, "Introducing the Knowledge Graph: things, not strings," Google blog, May 16, 2012 (https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html).

FORRESTER®
CHALLENGE THINKING. LEAD CHANGE.

We work with business and technology leaders to develop customer-obsessed strategies that drive growth.

PRODUCTS AND SERVICES

- Core research and tools
- Data and analytics
- Peer collaboration
- Analyst engagement
- Consulting
- Events

Forrester's research and insights are tailored to your role and critical business initiatives.

ROLES WE SERVE

Marketing & Strategy Professionals

CMO
B2B Marketing
B2C Marketing
Customer Experience
Customer Insights
eBusiness & Channel Strategy

Technology Management Professionals

CIO
- Application Development & Delivery

Enterprise Architecture
Infrastructure & Operations
Security & Risk
Sourcing & Vendor Management

Technology Industry Professionals

Analyst Relations

CLIENT SUPPORT

For information on hard-copy or electronic reprints, please contact Client Support at +1 866-367-7378, +1 617-613-5730, or clientsupport@forrester.com. We offer quantity discounts and special pricing for academic and nonprofit institutions.