Catching Fire: Start-Ups in the Text Analytics Software Industry

Aaron W. Baur
ESCP Europe Business School Berlin
abaur@escpeurope.eu

Max Breitsprecher
ESCP Europe Business School Berlin
mbreitsprecher@escpeurope.eu

Markus Bick
ESCP Europe Business School Berlin
mbick@escpeurope.eu

Abstract

While there is a certain amount of M&A activity in the software industry, it is far from consolidating. Start-ups in certain parts of this dynamic industry continue to thrive. In the case of Business Intelligence and Business Analytics (BI&A), more and more players enter the stage and either offer full packages or claim niche positions. The BI&A sub-market of Text Analytics is well researched regarding technical basics and dimensions. However, a comprehensive investigation of drivers and dynamics that encourage entrepreneurs to start their businesses in this rapidly growing market is missing. This research-in-progress paper provides an understanding of developments in this flourishing industry sector. After an extensive market study combined with theoretical, literature-based reasoning, this paper presents seven propositions that will be analyzed and tested with qualitative expert interviews and a large-scale quantitative survey in a next step.

Keywords

Entrepreneurship, Start-Ups, Software, M&A, Text Analytics, Business Intelligence, Business Analytics

Introduction

In its July 2013 Hype Cycle for Emerging Technologies, Gartner reports Big Data as being just before and Content Analytics being just after the "Peak of Inflated Expectations" (Gartner 2013). Both terms can be classified under the umbrella term Business Intelligence and Analytics (BI&A), which in a more practical language may be regarded as tools and processes handling that “Big Data”. Important IS & CS related journals like MIS Quarterly, Decision Support Systems or Communications of the AIS frequently publish articles related to that domain (Chen et al. 2012). Moreover, other scientific disciplines see Big Data as a key theme now and in the future. For instance, the Marketing Science Institute calls it one of their seven Research Priorities in 2012-2014 (MSI 2012). Additionally, the popular press is full of Big Data articles and reports, highlighting its potential for businesses and the economy as a whole (Economist Intelligence Unit 2012, Financial Times 2013, New York Times 2012). But not all fields possibly attributable to Big Data develop at the same pace.

Here, we suggest having a closer look at the Text Analytics (TA) software market to shed light on its increased relevance in terms of an entrepreneur’s attractiveness. Seven research propositions try to investigate the “how”, “where” and “why” of economic activity in this sector. By testing these propositions, the research adds to the discussion in three ways: It advances academia by elaborating the knowledge of drivers in disruptive software markets. It helps practitioners in TA or related fields to get a feeling for the speed, particularities and advances in this market. And it informs policymakers in markets where a high TA start-up activity can be seen (e.g., US, EU) and where it is missing (e.g., Africa, Asia). The latter ones can then work out policy changes that stipulate more local activity in this sector.
Relevance of Text Analytics

The Text Analytics industry develops along a rapidly expanding “digital universe”. While data can be generally stored in three different forms – structured (e.g., numbers in a database), semi-structured (e.g., text in an Excel sheet) and unstructured (e.g., product review on a homepage, email message, blog entry) – the clear majority of data is found in the semi- and unstructured arrangement, making up between 80%-95% of all data (see Table 4). Until recently, it had been a time and resource consuming process to carry out reliable and decisive insights from these information sources. Text Analytics software offers the possibility to automate this process of unstructured data extraction using sophisticated algorithms helping to save time, costs and broaden the scope and depth of the analysis.

Prior research has often focused either on technical aspects of Text Mining and Text Analytics (Chaudhuri et al. 2011) or is concerned with applying the technology effectively in a predefined domain or for a specified task. Published research includes studies on health-care (symptom discovery and management (Chen 2011, Wactlar et al. 2011)), law enforcement, financial trading and securities fraud detection (Abbasi et al. 2012), intelligence services and counter-terrorism (Chen et al. 2008, Chen 2006), e-government and politics (Yang and Callan 2009), market intelligence (Doan et al. 2011) and many other topics.

At the same time, authors argue that most of the future applications are still to be discovered – affecting basically every sector of our economy (Reamy 2012). In addition, the combination of existing tasks and the integration with other technologies (e.g., machine learning) will offer great opportunities for businesses, organizations and individuals in the future, classifying Text Analytics as a potentially disrupting technology.

For years, the use of Text Analytics remained a niche market developing under the wider area of Business Intelligence and Analytics. Only recently have market research companies like Gartner and large enterprises recognized the industry as an independent line of business (Badenhorst and Fitzgerald 2013). Due to the fragmented start-up nature of the TA industry, most of the rather small companies are privately held and therefore do not disclose account statements, making exact estimations of market size and growth rates difficult. However, some estimates are published by analysts and displayed in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>2015</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Business Intelligence</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Market Size (million USD)</td>
<td>10,546</td>
<td>12,295</td>
<td>13,131</td>
<td>14,079</td>
<td>15,122</td>
<td>16,223</td>
</tr>
<tr>
<td>Market Growth (p.a.)</td>
<td>13.4%</td>
<td>17.1%</td>
<td>6.8%</td>
<td>7.2%</td>
<td>7.4%</td>
<td>7.2%</td>
</tr>
<tr>
<td><strong>Text Analytics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Market Size (million USD)</td>
<td>835</td>
<td>1,044</td>
<td>1,305</td>
<td>1,631</td>
<td>2,039</td>
<td>2,548</td>
</tr>
<tr>
<td>Market Growth (p.a.)</td>
<td>25.9%</td>
<td>25%</td>
<td>25%</td>
<td>25%</td>
<td>25%</td>
<td>25%</td>
</tr>
</tbody>
</table>

Table 1. Market Size and Growth Estimations of BI and TA Market

The general Business Intelligence market is expected to grow around 7% year-on-year from 2012. The growth of the TA market, albeit starting from a much lower basis, is anticipated by analysts to account to at least 25% until 2015.

Entrepreneurship in BI&A and Text Analytics

Of course, the promising growth figures and the increased significance in both practice and research has not gone unnoticed from tech savvy business graduates and software engineers thinking about starting their own business. Many entrepreneurs focus on the area of Text Analytics in particular to establish their

---

new ventures. However, few studies were conducted highlighting the increased importance of the TA industry as a growth factor for knowledge-based economies. Grimes (2005) was one of the first to pursue a market segmentation, describing go-to-market strategies and emphasizing business strategies, new players and synergy potentials. Building on his work, Grimes (2009a, 2011) conducted a comprehensive vendor survey, revealing market drivers, opportunities and threats as well as estimating market size and growth. In addition, Halper et al. (2013) provide an overview of market trends and rank leading software firms in terms of the companies’ strategy, vision and strength as well as product offerings. Commercial market reports were, inter alia, produced by DMG Consulting (2010), Yuen and Koehler-Kruener (2012) and Technavio (2012).

Furthermore, some important industry aspects have not been studied until now, including its start-up dynamics and Merger & Acquisition (M&A) activities. Harnessing data form a comprehensive market analysis, the increased attractiveness and relevance of TA for entrepreneurs will be discussed in the course of this paper.

**Term Definition and Common Capabilities**

Although researched in the scientific literature for some decades, no commonly shared definition for the term Text Analytics has emerged. Prior to researching the industry and its appeal to entrepreneurs, however, a definition able to confine and classify market participants has to be found.

Drawing up on various definition attempts of Feldman and Dagan (1995), Hearst (2003), Grimes (2005, 2007b), IBM (2010) and Halper et al. (2013), the following definition shall be used for this work: *Text Analytics is a technology and process to discover new information and value in a large corpus of linguistic entities, by (at least) semi-automatically extracting and structuring information from weakly structured sources.*

Commonly used synonyms for Text Analytics include Text Mining, Text Data Mining, Text Knowledge Engineering, Knowledge Discovery in Text, Intelligent Text Analysis and Text ETL. While those terms are used interchangeably by some authors, others clearly define them as separate domains (Miner 2012). In the presented paper only the term Text Analytics or its acronym TA will be used since it developed itself to be the most commonly used in the business domain (Fan et al. 2006).

Common capabilities of TA that explain its relevance in a more and more data-driven world include categorization and clustering (Hung 2012), information extraction, sentiment and opinion analysis (Aggarwal and Zhai 2012, IBM 2010, Miner 2012, Pang and Lee 2008, Reamy 2012), concept linkage (Fan et al. 2006, Zhong et al. 2012), extractive and abstractive summarization (Aggarwal and Zhai 2012), topic tracking (Rajamaran and Tan 2001, Rajman and Besancon 1998) and question answering (firstly developed by the Massachusetts Institute of Technology, see Katz et al. 2007, Maybury 2004). Especially advances in sentiment/opinion analysis and information extraction are regarded as key capabilities that instigate growth of importance of TA (Reamy 2012).

The remainder of this paper is structured as follows: After clarifying the applied methodology, seven research propositions regarding the relevance of TA to business ventures are pronounced and first insights discussed. The paper closes with conclusions and an outlook on next steps planned to complete the research.

**Research Methodology**

This proposed study will be conducted using a mixed-methods design, which has been found appropriate for connecting exploratory with confirmatory research (Creswell and Plano Clark 2010).

First, a complete database search has been executed to receive a comprehensive global market study identifying players in the Text Analytics software industry. Sources used were company databases (Hoovers, Orbis, D&B, LinkedIn, Crunchbase, AngelList), industry publications, market research reports, and company websites.

Second, to discern potential drivers regarding the observed increased start-up activity in the Text Analytics industry, a first exploratory phase has been started. The literature study will be conducted using selected large-scale and reputable digital libraries in IT, engineering and related fields. Additionally,
industry/trade magazines will be screened electronically. To remain contemporaneous, only articles dating back to 2009 will be used. The sources are Web of Science, EBSCO Business Source Complete, IEEE Xplore, ScienceDirect, and the ACM Digital Library (cf. Chen et al. 2012). Search terms are comprised of a variety of queries including “text analytics/mining market”, “text analytics/mining industry” etc. in conjunction with “entrepreneurship”, “start-up”, “business founding”, “business ventures” and others. Third, besides the analysis of published literature, we have planned and already partially conducted qualitative semi-structured expert interviews (Hair et al. 2011). Interviewees include software industry experts, venture capitalists and TA entrepreneurs. Entrepreneurs in this respect are defined as founders and co-founders of firms that were legally established within the last three years of this study. The three-years’ time span was set as to make sure the entrepreneur can still correctly recall the decisive drivers that led him/her to start the business, as these can change during the course of business activity. The interviews will be transcribed and coded using MAXQDA software. After reaching a theoretic saturation, i.e., the state that few or none new driving forces come up, the exploratory phase ends.

Fourth, based on the qualitative interviews, a second confirmatory phase will follow. Hence, a survey will be designed and distributed to the TA companies detected in the market analysis. Again, only “young” companies (i.e., maximum of three years since establishment) will enter the sample.

The combination of these four research methods will ensure insightful and reliable results.

**Research Propositions**

Various factors may contribute to an increased start-up activity in the TA software market. In the following section, we discuss seven research propositions that we judge worthy of exploring.

**Number, Founding Date and Global Diversity of Players**

The numerous Web 2.0 applications developed and commercialized in the second half of the 2000s created an abundance of user-generated content from different online social media such as (micro) web logs (blogs), social networking sites (incl. videos and photos), forums, chats, online games and online groups (Lusch et al. 2010). With these developments, the digitalization of content increased exponentially (Aggarwal and Zhai 2012). In addition, companies’ internal communication and processes as well as customer and supplier engagement became increasingly digital (e.g., paperless office). Especially the desire to understand customer thoughts, opinions and remarks fueled the development of the Text Analytics industry (Halper et al. 2013).

The market study identified and categorized 372 companies that develop and sell Text Analytics software. It has to be noted that not all companies have complete TA packages in their portfolio, but rather focus on certain modules and target specific industries. The dynamics offer roughly three waves of development.

![Figure 1. Number and Year of Establishment of Firms in the TA Industry](image-url)
First, between 1985 and 1998, only two to three companies were founded annually. Second, between 1999 and 2003, eleven companies were established on average per annum. The year 2004 marks an outlier with 5 companies. Between 2005 and 2012, more than 30 firms were founded in average, with a clearly increasing trend (see Figure 1).

From 2007/08 on, when social media and Web 2.0 effects of creating vast amounts of data fully took up speed, 224 firms were started in this technology domain. When looking at the countries of establishment, the USA and Canada stick out. Here, 97 companies were founded, representing about 43% of all firm setups in this period. All European countries taken together reached 90 company establishments, i.e., 40%. The rest of the world merely contributed 16.5% (see Table 2). Current numbers until September 2013 do not show any reversal of trend.2

In other words, there is a clear trend towards an increasing pace of company setups. However, although there is some start-up activity in Asia and Latin America, no development towards a declining supremacy of the US and Europe can be observed.

<table>
<thead>
<tr>
<th>Year</th>
<th>USA/Canada</th>
<th>Europe</th>
<th>Latin America</th>
<th>Asia</th>
<th>Australia/New Zealand</th>
<th>Africa</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>8</td>
<td>9</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td>2012</td>
<td>18</td>
<td>15</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>2011</td>
<td>19</td>
<td>21</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>43</td>
</tr>
<tr>
<td>2010</td>
<td>13</td>
<td>11</td>
<td>3</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>35</td>
</tr>
<tr>
<td>2009</td>
<td>10</td>
<td>15</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>29</td>
</tr>
<tr>
<td>2008</td>
<td>19</td>
<td>13</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>36</td>
</tr>
<tr>
<td>2007</td>
<td>10</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>2007-2013</td>
<td>97</td>
<td>90</td>
<td>13</td>
<td>20</td>
<td>4</td>
<td>0</td>
<td>224</td>
</tr>
<tr>
<td>% of total</td>
<td>43.3%</td>
<td>40.2%</td>
<td>5.8%</td>
<td>8.9%</td>
<td>1.8%</td>
<td>0.0%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 2. Number and Country of Establishment of Start-Ups in the TA Industry

We therefore claim the following proposition:

*Proposition 1: While the overall number of start-ups in the TA industry increases, there is currently no development towards a stronger international influence, i.e., the US dominance remains for the near future.*

**Oligopolistic Margins**

Industry analysts’ reports suggest that the top Text Analytics vendors, measured by revenue, claim a prominent position and can be regarded as the “Big Four”. These are HP Autonomy, IBM, SAP, and SAS (see Table 3). How much of their revenue can be assigned specifically to TA remains unclear, as the published accounts do not disclose numbers for each area of business, i.e., the numbers are aggregated and do not allow a distinct business unit related separation. Several interviewed experts commonly agreed in their estimates that these four firms together represent more than 50% of the total TA industry revenue.

---

2 The numbers in 2013 only represent the months January to September.
Table 3. The “Big Four” in the TA Software Market

<table>
<thead>
<tr>
<th>Company Name</th>
<th>Main Text Analytics Software Solution(s)</th>
<th>Headquarters</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP Autonomy</td>
<td>HP IDOL</td>
<td>USA</td>
</tr>
<tr>
<td>IBM</td>
<td>SPSS Modeler, IBM Content Analytics (LanguageWare)</td>
<td>USA</td>
</tr>
<tr>
<td>SAP</td>
<td>SAP HANA, SAP Business Suite</td>
<td>Germany</td>
</tr>
<tr>
<td>SAS</td>
<td>SAS Text Analytics, Enterprise Miner</td>
<td>USA</td>
</tr>
</tbody>
</table>

According to market reports, the primary contenders of the top four vendors include Attensity, Basis Technology, Clarabridge, Clearforest, Leximancer, Megaputer, OpenText and Temis. The companies generate annual revenues between 10-40 million USD with their Text Analytics business. Those contenders are either specialized in Text Analytics solutions or they take at least a more prominent role in their portfolio. Experts estimate the sum of their revenue to account for more than a quarter of global Text Analytics revenues. Hence, less than 25% of revenue can be assigned to the approximately 360 remaining smaller enterprises and start-ups in the domain.

Every month, Damodaran (2014) from NYU Stern tracks and calculates profit margins of 96 industries. Data from January 2014 indicates that the 273 players in the “Computer Software Industry” have an average net margin of 20.47% (26.73% pre-tax unadjusted operating margin). With this being the eighth highest margin of the 96 industries considered, software firms in general earn a higher-than-average profit. There are no specific numbers for the TA business units of the Big Four available. However, we reckon that profits in the TA segment are at least as high as this overall average, as more commodity-like software products (e.g., Enterprise Resource Planning or Operating Systems) are also included in this margin. With the odd-shaped distribution of current market participants, i.e., four firms contributing 50% of revenue, eight firms additional 25% and hundreds of other firms the rest, it is reasonable to assume that high oligopolistic profit margins are noticeable and will continue to lure new players into the market.

**Proposition 2:** The odd-shaped Text Analytics market yields oligopolistic margins that contributes to increased start-up activity in the Text Analytics software industry.

**Mergers & Acquisitions**

There is much literature on the drivers of M&A in high-tech industries in general (see Rossi et al. 2013 for an overview) and some research of M&A in the software industry (Laamanen et al. 2013, Singh et al. 2013, Kakola 2003). Frequently, sector consolidation is attributed to an industry becoming mature and remaining players seeking higher growth rates. However, we want to argue that, although M&A activity is happening, the TA market is a good example of one fraction of the software industry that still does not consolidate, but rather becomes more fragmented.

In two different studies, Gao and Iyer (2006, 2008) researched whether M&As (2006) and alliances (2008) of companies in the software industry were more successful if they are more complementary to each other. In both studies, this was the case, arguing that it certainly does make sense for players like SAP or IBM to write a shopping list and acquire smaller related niche players in order to round off their product portfolio with supplementary applications and technologies (Cho et al. 2013).

Unified procurement, i.e., customers having a “one stop shop” for their software needs, is also mentioned as a driver for consolidation (Kauffman and Tsai 2009). The concept of “relatedness” is researched by Tanriverdi and Lee (2008) who claim that software firms investing in within-industry competitors achieve higher profit margins. Relatedness can be software-platform related or product-market related.

In order to monitor developments of the TA industry, a thorough analysis of M&A deals was carried out. In addition to the sources mentioned in the methodology part, Bureau van Dijk’s Zephyr Database, which is regarded as the world’s most extensive database on M&A, was used. In the time span between 2000 and 2013, 49 M&A deals could be identified, yielding three important aspects:³

---

³ The data can be requested from the authors.
First, major multinational technology companies acquired Text Analytics start-ups over time, strengthening their already existing in-house efforts and capabilities in this domain or acquiring new capabilities. All of the four identified largest TA vendors showed M&A activities in the past. For example, Reuters Group, the British multinational media and financial information company, was one of the first to acquire a conventional Text Analytics vendor, Clearforest, in 2007. SAP acquired Business Objects in the same year, which paved the way for them into BI in general and TA in particular. In 2008, SAS bought Teragram and Microsoft entered the market with its acquisition of Fast Search&Transfer. In 2009, IBM acquired SPSS, Oracle bought Endeca Technologies Inc. in 2011. In the same year, Hewlett-Packard (HP) spent an estimated 10.2 billion USD to acquire the content analytics provider Autonomy, which is by far the largest deal to date. In 2013, Yahoo!, Google and Twitter extended their existing Text Analytics capabilities by acquiring smaller technology start-ups. As these examples indicate, corporations either buy companies to strengthen their presence as Text Analytics solution providers or to improve their core business functions and products through the integration of Text Analytics technology without offering TA products and services directly.

Second, it becomes clear from the dataset that there are significant differences in the M&A strategy depending on the country of origin of buyer and target. In the 49 mergers and acquisitions undertaken over time, a majority of 29 was executed by US companies, followed by British (8) companies as acquirers. US companies therefore pursue an aggressive M&A strategy, showing strong activities as acquirer globally. European companies on the other hand, more reluctant in their M&A activities, buy significantly more often companies located in other European Union countries than outside the EU. Reasons can be multifold, including a more common business culture of acquisitions as growth strategy, higher availability of risk capital and a stronger technology sector. The finding that UK companies are second in acquiring other enterprises in this industry supports those considerations.

Third, no notion of an accelerating consolidation of the industry due to M&A can be observed. Neither the number nor the value of transactions is increasing significantly. In most cases, companies seem to selectively buy companies with capabilities and technologies in related fields of their core business, instead of pushing for consolidation and acquiring direct competitors. Taking the recent deals undertaken by Yahoo! and Google, this point can be exemplified. Yahoo! and Google as major web search providers already possess extensive Text Analytics capabilities. Nevertheless, they acquired Summly and Wavii, both capable of generating automatic summaries of text entities, to improve their search offers. This missing consolidation of the market through M&A activities could be triggered by the prevailing low entry barriers. These low entry barriers provoke a continuous and increasing stream of new market entrants, making the acquisition of competitors out of pure consolidation efforts futile.

Proposition 3: Although merger & acquisition deals are existent in the TA market, the industry is not yet on a route to consolidation, indicating continued potential for start-up activity in the Text Analytics software industry.

Increased Demand and Less-than-Satisfied Customers

The raw material for Text Analytics applications consists of massive volumes of unstructured textual data. The growth of the “digital universe” accelerated at an unprecedented speed. Gantz and Reinsel (2012) estimate that the amount of data stored globally will increase from 1,227 exabytes in 2010 to almost 7,910 exabytes in 2015 – a compound annual growth rate of 45%. Data generated by enterprises is estimated to grow even faster by a rate of 50% annually, from around 200 exabytes (2010) to over 1,600 exabytes (2015). At the same time, merely a small fraction of this pool has been explored for analytic value, although analysts estimate that by 2020, possibly a third of the internet will contain information that might be valuable for businesses if properly analyzed.

However, today TA still only captures a niche of the much larger Business Intelligence software market (see Table 1). Almost all enterprise activities were originally targeted towards the “low hanging fruit of transactional data [...]. It was a quick win to transfer spreadsheets of information in neat columns and rows” (Taylor, 2013). Hence, analyzing structured data became common in the last several years, whereas analyzing semi- or unstructured data still remains more the exception to the rule. From a return on

\[4\] The dates represent the time of deal announcement.
Investment (ROI) view, this did make sense in the past, since analyzing structured data in relational databases is technologically easier than processing large amounts of unstructured data, and the necessary tools were widely available.

However, now, the bulk of data stored in general and by enterprises in particular remains unstructured data (see Figure 2).

![Figure 2: Breakdown of Globally Stored Data by Type (in Exabytes)](image)

Estimates of stored unstructured data range from 31% to over 90% of data (see Table 4). Nair and Narayanan (2012) estimate the unstructured data growth being 15 times the rate of structured data growth. IDC states a CAGR of 56% for unstructured data volume, compared with 20% CAGR for structured data volume.

<table>
<thead>
<tr>
<th>Source</th>
<th>Year</th>
<th>Scope</th>
<th>Share of Data Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM</td>
<td>2006</td>
<td>All data</td>
<td>80% unstructured</td>
</tr>
<tr>
<td>Russom (tdwi)</td>
<td>2007</td>
<td>Enterprise data</td>
<td>31% unstructured, 47% structured, 22% semi-structured</td>
</tr>
<tr>
<td>SAS Institute</td>
<td>2009</td>
<td>Enterprise data</td>
<td>70% unstructured, 25% structured, 5% semi-structured</td>
</tr>
<tr>
<td>DMG</td>
<td>2010</td>
<td>All data</td>
<td>90% unstructured</td>
</tr>
<tr>
<td>IBM Research</td>
<td>2011</td>
<td>All data</td>
<td>85% unstructured</td>
</tr>
<tr>
<td>IDC</td>
<td>2012</td>
<td>All data</td>
<td>85% unstructured</td>
</tr>
</tbody>
</table>

Table 4. Estimates for share of unstructured to structured data

This sheer of amount of data to be analyzed will create demand for software solutions. This will partially be satisfied by current software suppliers, but creates room for additional market entries.

Another facet that contributes to increased demand is customer satisfaction. Research by Miles (2012), surveying 402 AIIM members, indicates that 74% of respondents would find it valuable or very valuable to carry out sophisticated analytics on unstructured content. However, another study reports an imbalance in terms of products and associated service. While 70% of TA users are completely satisfied or satisfied with the technology as such, only 42% were satisfied with the availability of professional services and support. This results in only 49% of users who would generally recommend their current Text Analytics vendor and 28% who would actively do so. In addition, a majority of current users reports service aspects to be of primary importance, not technical features (Grimes 2011a).

We hence infer the rapidly increasing number of first time customers along with a high number of dissatisfied current Text Analytics users will create solid business opportunities for emerging entrepreneurs, especially concerning more service-oriented offerings.

**Proposition 4:** The hugely increasing amount of unstructured data and partially dissatisfied customers initiate demand for new offers that will increase start-up activity in the Text Analytics software industry.

---

5 Based on Gantz and Reinsel 2012.

Increased Hardware Capacities

The advances in processing power, memory and data storage have been significant over the last decades, making TA applications at all possible. The cost of creating, capturing, managing and storing information is down to 1/6 of the cost in 2005 and it continues to decrease quickly (Gantz and Reinsel 2012). In addition, “new data storage platforms (NoSQL, Hadoop, SAP HANA) condition the take-up of big data and large-scale analytics applications” (LT-Innovate 2013, p. 21).

These advances in hardware infrastructure now enable real-time Text Analytics applications, previously impossible to realize out of a lack of affordable hardware. We argue that entrepreneurs will take advantage of this development and will try to make up leeway, i.e., start new business ventures.

**Proposition 5:** Technological advances in data storage platforms and reduced costs will increase start-up activity in the Text Analytics software industry.

Advances in Structured Data Analytics

It is fair to say that structured data analytics paved and still paves the way for Text Analytics adoption. Although Text Analytics was conceived earlier as a main Business Intelligence technology, because of its higher complexity its adoption and technological advances progressed slowly at first. Structured data analytics, however, were easier to apply, offering higher precision levels at lower costs, which in turn offered the possibility of higher ROI rates. With time, considerable investments were put into structured data analytics projects; processes, techniques, visualizations and algorithms were improved and the “low hanging fruits” of Business Intelligence and Analytics were harvested. During that process, showcase applications emerged and user expectations as well as an understanding formed in organizations, that value can be extracted from increasing volumes of data. In its 2013 survey of 2,000 CIOs, Gartner revealed that Analytics and Business Intelligence is the top technological priority in enterprises. In addition, the increasing interest in leveraging various sources of data can be seen in the hype around “Big Data”, starting in 2011. Currently, an increasing priority in leading enterprises is the integration of structured and unstructured data sources, fueling the adoption of Text Analytics (Halper et al. 2013).

We argue that the fast advances in structured data analytics also accelerate developments in the semi- or unstructured data analytics realm.

**Proposition 6:** Advances in structured data analytics will support developments in the unstructured data analytics area and thus increase start-up activity in the Text Analytics software industry.

Increased Attention in Education, the Open-Source Community & Venture Capital

In their article “Business Intelligence and Analytics: From Big Data to Big Impact”, Chen et al. (2012, p. 1166) quote an interview with Hal Varian, Chief Economist at Google and former professor at Berkeley, “So what’s getting ubiquitous and cheap? Data. And what is complementary to data? Analysis. So my recommendation is to take lots of courses about how to manipulate and analyze data: databases, machine learning, econometrics, statistics, visualization, and so on.” Taking this advice seriously, many universities and colleges are setting up degree programs in BI&A and Text Analytics, e.g., the University of Texas at Arlington7 or the University of Illinois at Urbana-Champaign8. Likewise, the UK established the first publicly founded TA center worldwide – the National Centre for Text Mining (NaCTeM), and the European Commission promotes consortium and university initiatives throughout the continent9.

In these programs, not only core TA skills like Statistics and Database Science are taught, but also related technologies trained. This is important, since TA borrows many of its characteristics, techniques, procedures and algorithms from related fields, notably Data Mining (Manyika et al. 2011), Natural Language Processing (Kao and Poteet, 2007), Content Management, Information Retrieval and Corpus-

---

7 http://www.utexas.edu/insyopma/
8 http://www.stat.illinois.edu/degrees/msanalytics.shtml
9 For a detailed list of European academic initiatives see Grimes 2007c.
based Computational Linguistics (Hearst 1999), Artificial Intelligence and Machine Learning (Miner 2012). Hence, graduates in computer science, informatics, statistics and business information systems find more and more highly relevant educational offers. With the knowledge gained from these classes, a first but important step to becoming a TA entrepreneur has been taken. Additionally, an ecosystem of regular expert events, industry associations, networks\textsuperscript{10}, research communities and partnerships has started to establish itself around the evolving Text Analytics domain. With the Text Analytics Summit, Text Analytics World and Sentiment Analysis Symposium, three regular conferences are held annually next to the dozens of vendor conferences detained worldwide. Second, there is a plethora of open source tools and systems available on the web. A non-exhaustive list of open-source projects and solutions is displayed in Table 5.\textsuperscript{11} These open-source tools generally provide core services and can be customized or extended to provide more advanced or more-specialized capabilities.

\begin{table}[h]
\centering
\begin{tabular}{|l|l|}
\hline
\textbf{Organization/ Project/ Author} & \textbf{Open-Source Software} \\
\hline
Alias-i & LingPipe \\
\hline
Apache Software Foundation & OpenNLP (Diverse Software Solutions) \\
\hline
Apache UIMA Project & UIMA \\
\hline
Carrot\textsuperscript{2}-Project & Carrot\textsuperscript{2} \\
\hline
Datumbox & Datumbox API \\
\hline
Higuchi, Koichi & KH Coder \\
\hline
Fraunhofer Institute for Technological Trend Analysis & Idea Web Miner, Technological Idea Miner, Idea Patent Miner, others \\
\hline
IKANOW & IKANOW Infinit.e \\
\hline
KNIMEtech & KNIME Text Processing \\
\hline
KnowItAll Project & ReVerb \\
\hline
Liu, Bing; Li, Xiaoli & S-EM (Spy-EM), LPU \\
\hline
NLTK-Project & Natural Language Toolkit (NLTK) \\
\hline
Padró, Lluís & FreeLing \\
\hline
R Development Core Team & R – tm (Text Mining Package) \\
\hline
Rapid Miner & RapidMiner Text Mining \\
\hline
Redonnet, Jean-Pierre & TexLeyAn \\
\hline
Rehurek, Radim & Gensim \\
\hline
Semantic Indexing Project & The Semantic Engine \\
\hline
Social Science Consulting & TextQuest \\
\hline
Tagul & Tagul \\
\hline
Thomson Reuters & Open Calais \\
\hline
University of Ljubljana & Orange-Text (Add-on to Orange suite) \\
\hline
University of Sheffield & GATE-Project \\
\hline
\end{tabular}
\caption{Open-Source Text Analytics Software}
\end{table}

\textsuperscript{10} For example, the LinkedIn group ‘Text Analytics – revealing unstructured data’ with more than 10,000 members may be the largest social network on TA worldwide.

\textsuperscript{11} This compilation is preliminary and therefore not exhaustive. It misses in particular application-specific open-source solutions, e.g., from biomedical science, see http://arrowsmith.psych.uic.edu/arrowsmith_uic/tools.html for an overview.
While these open-source solutions can constitute viable alternatives to commercial software suites (Sen et al. 2011), they also help start-ups and other solution providers to focus their developments on their specific competitive advantage, and on their customers’ real needs. A reinvention of the wheel is not necessary, as core modules may be suitable with moderate alterations. Additionally, industry partnerships with marketing services providers and database content providers further lower the entry barriers (Grimes 2010). Hence, starting your own business becomes easier.

Finally, venture capital (VC) firms are increasingly open to investments in BI&A and TA software and hardware companies. The VC industry noticed that more than 110 companies have entered the market in the past three years (see Figure 1) and invested accordingly. Between Q3/2012 and Q3/2013, 1.37 billion USD were invested in various data analytic companies in the US alone, which means an increase of 217% to the previous year (Datoo 2013). On average, 12 VC deals were done per quarter, while this development is estimated to accelerate (GP Bullhound 2013). For businesses to receive seed and risk capital funding for their ventures in the TA industry is therefore easier than ever.

Proposition 7: Increased number and community attention of TA relevant educational offerings, a high availability of open-source software and raised venture capital interest will increase start-up activity in the Text Analytics software industry.

Contribution, Conclusion and Outlook

This research-in-progress paper examines the Text Analytics software market and the drivers for continued high start-up activity within this field. It therefore contributes threefold: To academia by pointing out developments and particularities of TA as a young market with a disruptive character. To practitioners in TA or related fields in order to get a feeling for the speed and advances in this market. And to policymakers in markets where a high start-up activity can be seen (e.g., US, EU) and where it is missing (e.g., Africa, Asia). The latter ones may find helpful hints which levers to pull in order to encourage more local activity in this sector.

Within the study, seven research develop propositions based on past research, a market study and initial qualitative expert interviews were developed. As preliminary results suggests, driving forces of an increased start-up activity in the Text Analytics software market are oligopolistic margins, hugely increasing amount of unstructured data and partially dissatisfied customers, advances in data storage platforms and structured data analytics, an increased number and community attention of TA relevant educational offerings, a high availability of open-source software and raised venture capital interest. Some of these drivers apply to the software sector in general, others apply to the BI&A industry and still others are very specific to the TA market.

The propositions need to be thoroughly researched and tested. We suggest additional expert interviews as well as a large-scale quantitative survey among TA entrepreneurs, customers and industry experts. In the course of the project, complementary research methods may be necessary to further support the arguments.
REFERENCES


