FP7-ICT Strategic Targeted Research Project TrendMiner (No. 287863)
Large-scale, Cross-lingual Trend Mining and Summarisation of Real-time Media Streams

D8.2.1 Market Watch - v1

Paul Ringler (Editor, SORA), Mauro Navarra (EK), Francesca Spagnoli (EK), Thierry Declerck (DFKI), Renisha Chainani (EK), All (for Exploitation plans)

Abstract
FP7-ICT Strategic Targeted Research Project TrendMiner (No. 287863)
Market Watch v1 D8.2.1 (WP 8)

This deliverable gives a first outline of the commercial and scientific relevance of TrendMiner for the general public. We cover the ecosystem of Social Media which forms the technological and cultural environment for this project and provide an estimation of the market for commercial products based on this project. We outline the technological background of natural language processing and stream reasoning software, and give examples for existing providers of tools based on similar use cases. Finally we illustrate the commercial and scientific opportunities created by this project, as well as more concrete exploitation goals.

Keyword list: Social Media, Natural Language Processing, Open Source Software, Business Intelligence, Financial Use Case, Political Use Case

Nature: Report Dissemination: PU
Contractual date of delivery: 1.11.2012 Actual date of delivery: 5.11.2012
Reviewed By: IMR, DFKI
Web links: http://trendminer-project.eu/

CHANGES

Version	Date	Author	Changes
0.1	19.10.2012	Paul Ringler	Creation, established document structure, Entered available contributions, editing for content and style.
0.2	22.10.	Francesca Spagnoli and Mauro Navarra (EK)	Integration of scientific and commercial opportunities, exploitation. Financial Content Analytics Providers descriptions.
0.3	25.10.	SORA	Further editing after Telco
0.4	26.10.	EK	Editing + Profiles
0.5	27.10.	DFKI	Contributions + General editing
0.6	29.10.	SORA	Improved references, added table of abbreviations, more edits.
0.7	31.10.	IMR	Review comments by IMR
0.8	01.11.	DFKI	Editing, on the base of comments by IMR and others
0.9	02.11	EK	Editorial round
1.0	2.11.	SORA	Final Version

TrendMiner Consortium

This document is part of the TrendMiner research project (No. 287863), partially funded by the FP7-ICT Programme.

DFKI GmbH
Language Technology Lab
Stuhlsatzenhausweg 3
D-66123 Saarbrücken
Germany
Contact person: Thierry Declerck
E-mail: declerck@dfki.de

University of Southampton
Southampton SO17 1BJ
UK
Contact person: Mahensan Niranjan
E-mail: mn@ecs.soton.ac.uk

Internet Memory Research
45 ter rue de la Revolution
F-93100 Montreuil
France
Contact person: France Lafarges
E-mail: contact@internetmemory.org

Eurokleis S.R.L.
Via Giorgio Baglivi, 3
Roma RM
00161 Italy
Contact person: Francesco Bellini
E-mail: info@eurokleis.com

University of Sheffield
Department of Computer Science
Regent Court, 211 Portobello St.
Sheffield S1 4DP
UK
Tel: +44 114 222 1930
Fax: +44 114 222 1810
Contact person: Kalina Bontcheva
E-mail: K.Bontcheva@dcs.shef.ac.uk

Ontotext AD
Polygraphia Office Center fl.4,
47A Tsarigradsko Shosse,
Sofia 1504
Bulgaria
Contact person: Atanas Kiryakov
E-mail: naso@sirma.bg

SORA Ogris and Hofinger GmbH
Bennogasse 8/2/16
1080 Wien
Austria
Contact person: Christoph Hofinger
E-mail: ch@sora.at

Hardik Fintrade Pvt Ltd.
227, Shree Ram Cloth Market,
Opposite Manilal Mansion,
Revdi Bazar, Ahmedabad 380002
India
Contact person: Suresh Aswani
E-mail: m.aswani@hardikgroup.com

Executive Summary

Since their development in the first decade of 2000, Social Media and related Web 2.0 applications have seen exponential growth in both numbers of users and relevance. Large volumes of multilingual data have become available and with this, the need to access and interpret them in a timely and affordable manner has grown.

The TrendMiner project aims at delivering innovative, portable, open-source, real-time methods for cross-lingual mining and summarisation of large-scale stream media, which can serve this need. TrendMiner implements this goal by proposing an inter-disciplinary approach, combining deep linguistic methods from text processing, knowledge-based reasoning from web science, machine learning, etc. Scalability and affordability will be addressed through a cloud-based infrastructure for real-time text mining from stream media.

This deliverable first outlines the ecosystem of Social Media applications (Ch. 2), and sheds light on the nature and function of the data that TrendMiner-based applications can access. Twitter plays a central role in this deliverable, since it has become an important point of exchange to communicate news, opinions and information very quickly. This makes it a prime source to detect topic and sentiment trends in an influencial section of the public.

Next, an overview of the target market for services based on TrendMiner is presented (Ch. 3) by comparing it to the existing market for Business Intelligence (BI) and Corporate Performance Management (CPM) applications. Like TrendMiner, these tools structure and filter large volumes of data, and help analysts to discover trends very quickly.

The following section (Ch. 4) elaborates on users and functions of the basic tool that will be developed in TrendMiner and describes the technological background of the software and hardware infrastructure developed in TrendMiner.

Following this, some examples of existing tools that have similar functions and purposes to what is being developed in TrendMiner are presented (Ch. 5). The focus is on content analytical tools that access streams of Social Media, news and financial data, providing visualisations, in-depth search, analysis functions and/or facilitate communication with users of Social Media.

In Chapter 6, every partner of TrendMiner is offering their view on possible commercial and scientific opportunities given by their contribution to both the R&D work and the use cases. Research partners also compare TrendMiner goals with existing systems. The use case partners show how TrendMiner-based tools can potentially increase their productivity and competitiveness in the mid- to long-term, while scientific partners show how their research efforts in the context of TrendMiner can advance beyond the technological and scientific status quo.

Finally, Chapter 7, offers more concrete insight into the particular exploitation plans (both scientific and commercial) of each partner in the consortium, showing how TrendMiner helps them to establish new contacts and/or develop new services.

Contents

Table of Abbreviations

API Application Programming Interface
B2B Business-to-Business
BI Business Intelligence
CPM Corporate Performance Management
CRM Client Relations Management
C-SPARQL Continuous SPARQL
DBMS Data Base Managment System
GATE General Architecture for Text Engineering
GEMET General Multilingual Environmental Thesaurus
HLT Human Language Technology
IE Information Extraction
IGGSA Interest Group on German Sentiment Analysis
LD Linked Data
LOD Linked Open Data
MCA Media Content Analysis
NLP Natural Language Processing
OBIE Ontology-Based Information Extraction
OSN Online Social Networks
OSS Open Source Software
POS Part of Speech
PR Public Relations
RDF Resource Description Framework
REST Representational State Transfer
RSS Really Simple Syndication
SEC Security and Exchange Commission
SPARQL SPARQL Protocol and RDF Query Language
TRNA Thomson Reuters News Analytics
URI Uniform Resource Identifier
WP Workpackage

1 Relevance to TrendMiner

TrendMiner is a Research and Development (R&D) project, combining 3 academic and research institutions and 5 SMEs interested in including innovative technologies in their product and services. In the course of TrendMiner, results are validated in two high-profile case studies: Financial decision support, with analysts, traders, regulators, and economists as possible target group, and political analysis and monitoring, with politicians, economists, and political journalists. It is very essential in those fields to have a real time access to streaming media. Since TrendMiner will not propose a purely research platform, but also economically viable solutions, this market watch is a pre-requisite in order to position the use cases of TrendMiner. While TrendMiner is not exclusively dealing with Social Media, but with different types of streaming data, the first version of this deliverable is focusing on this topic.

2 Background

2.1 Social Media and its Evolution

Social media is a family of online applications which enables users personifying different roles (customers, employees, consumers, companies, institutions, etc.) to share, co-create, discuss and modify user-generated content. This can happen on a variety of platforms, most of them are web based, but in recent times also mobile devices based. In other words, the interesting novelty in Social Media services is the fact the user is at the same time the data provider and no longer only an expert (writer) in a specific field. In this context, we see often the term ”user generated content”. The relevant Wikipedia article states here that ”The advent of user-generated content marked a shift among media organizations from creating online content to providing facilities for amateurs to publish their own content.”[1]

2.1.1 A short History of Social Media

These Social Media technologies were developed few years back – the oldest being blogs during 1998-1999, LinkedIn, Myspace 2003, Facebook in 2004 and Twitter in 2006. Late 2009 and 2010 however, were the years when Social Media gained true popularity and respect among consumers, brands and institutions. Twitter became a platform for breaking news and keeping the social world updated anywhere and at anytime. The number of Facebook users grew rapidly and so did for LinkedIn, Instagram, YouTube and etc. The consensus among many experts nowadays is, that Social Media cannot be ignored, when considering matters of sentiment or communication within a wider public.

Today Facebook has more than 1 billion users and this growth reaches beyond internet-enabled computers, as smartphone proliferation creates hundreds of millions of mobile consumers that make social networks their on-the-go digital portal. The shift from desktop access to mobile has not only increased the flow of information, but made it more real-time based.

2.1.2 A Types of Social Media[2]

The Social Media technologies encompass Internet forums, blogs, wikis, micro blogging, social networking sites (Twitter, Facebook, LinkedIn to name the most visited) and platform for photos and videos (Flickr, YouTube). The usage of Social Media is not restricted to personal activities and networking, but also includes professional information and connections. Companies and institutions may connect with their current and potential employees, Journalists, Politicians, and engaged citizens may exchange information, professional traders may use them to gain insights into possible market moving events.

[1] http://en.wikipedia.org/wiki/User-generated_content (last access 29.10.2012)

[2] http://www.mediassociaux.fr/2011/02/06/description-des-differents-types-de-medias-sociaux/ (last access: 29.10.2012)

Figure 1: Type of Social MediaTypes of Social Media[3]

In more detail, we can distinguish several types of Social Media:

Forum: A public discussion space where messages are displayed in chronological order.

Blog: A simplified publishing tool where items are displayed in chronological order and sorted into categories. Readers may submit comments that are moderated afterwards. The RSS feed allows easy export of content to aggregators and readers.

Wiki: An online knowledge base where users themselves write and edit articles on a any number of subjects.

Microblog: Service publishing, sharing and discussion based on very short messages. Accessible on the web, mobile devices, or through applications. E.g.: Twitter, Google Buzz, WhatsApp, etc.

Social Network: Site with restricted access where each user has a profile. Members are linked bilaterally or through groups. Some networks also offer more sophisticated features (messaging, publication and sharing content) and the ability to host third-party applications (platform). E.g.: Facebook, Xing, LinkedIn, Orkut, etc.

P2P Sharing: Online service where users can post pictures, videos, links and etc. Each element is linked to a published member and can be commented and rated. The notorious "Pirate Bay" service is among these services, but also other sites like YouTube, flickr, etc.

Aggregator: Online service to bring together all the publications of a user of Social Media (social stream). E.g.: FriendFeed, Hootsuite, etc.

Social gaming: Online games based on a social platform operator member profiles to offer different social interactions between players. E.g. FarmVille

Geolocation service: Applications to publish share and discuss over mobile devices. Articles or photos published are attached to a place to give them a geographic context. E.g. 4Square

[3] http://www.mediassociaux.fr/2011/02/06/description-des-differents-types-de-medias-sociaux/ (last access: 29.10.2012)

2.1.3 Characteristics of Social media

Social media have specific characteristics which differentiates it from traditional media (Mislove et al 2007):

User Based: The norm before the Social Media was that the content of the web pages was uploaded by a single entity (editor, company, etc.) and was read by other users on the Web. The flow of information was taken to be unidirectional, much like classical print media. By contrast, content on Social Media is created by the users, with minimal intervention from a moderator. The information can come from anyone who wants to participate. This combination of speed and variety makes Social Media attractive and useful for the users.

Interactive: Many Social Media are essentially applications that enable users to interact with each other. This is quickly becoming a pastime that more people are choosing over television – because it's more than just entertainment, it's a way to connect and have fun with friends.

Community-Driven: Social networks are built and thrive from community concepts. They provide virtual groups for people who share common beliefs or hobbies and allow them to share their views and interact with people having similar views.

Relationships/Connections: Unlike websites of the past, social networks thrive on relationships. The more relationships that you have within the network, the more established you are toward the center of that network.

Emotion over content: While traditional websites where focused on providing content, Social Media provides people with emotional support since they have their friends within easy reach.

Unorganized content: Due to the fact that the information of the Social Media is added by any user, this content is highly unorganized and unstructured as compared to any website of the traditional web. Information is often incomplete, sentences and words often truncated, and posted out of context.

Addictive: The major reason for the success of the Social Media is its addictiveness due to some of the above mentioned characteristics. According to estimates each user spends on an average 7 hours a month on Facebook. Compared to traditional media, this characteristic brings about many repeats, redirections, which distort the concept of weak/strong signals.

2.1.4 The Future

Social media will continue to grow with more people starting to use it. The future will see Social Media being incorporated into all areas of business from traditional CRM (Client Relations Management) to sales and marketing, with the intention to create a seamless connection to consumers. Technologies and tools will advance. ocial media raise also new challenges, based on its exponential growth: how to deal with this amount of information? How to build a bridge between Quantity and Quality?

2.2 The Social Media Ecosystem

The past years have seen a rapid growth of Social Media offers on the Internet. Still, a visible structure has emerged. Cavazza[4] currently sees three social networks as dominating the Social Media landscape: Facebook, Twitter and Google+. The large number of users they have attracted and the nature of services and applications they offer, suggests that this triad could remain stable for some time, as described by Cavazza: "I don't believe one can eat the two others, since each one have a distinct orientation: Twitter for content discovery, Google+ to manage your online identity and Facebook to interact with your friends".

It is interesting to note that a diverse and dense ecosystem of applications and social networks has come into being around these three main networks. Most of them fill the specialised niches left out by their larger peers, such as publishing, playing, buying, localisation etc. A good number of them offer links to other networks, especially the three largest one (in order to expand their own reach). For example it is now possible to write a blog entry on a WordPress[5] account and advertise this activity via Twitter. Similarly, someone may read that blog and post the link and a comment on Facebook.

Fred Cavazza proposes a graphics for representing the "Social Media landscape"(see Figure 2):

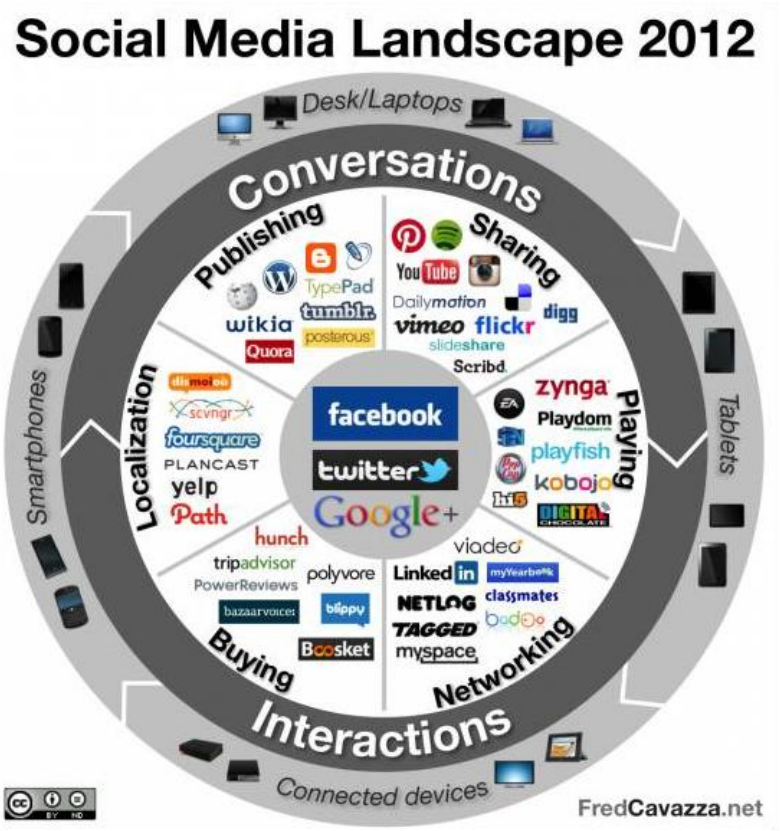

Figure 2: Social Media Landscape in 2012[6]

[4] http://www.forbes.com/sites/fredcavazza/2012/03/12/an-overview-of-the-social-media-ecosystem/ (last access: 29.10.2012)

[5] http://**wordpress**.com

[6] Image taken from: http://blogs-images.forbes.com/fredcavazza/files/2012/03/Social_Media_Landscape_2012.png (under Creative Commons Licence)

In addition to social networks that fill uncovered niches, another aspect of this ecosystem is the development of applications that thrive off the substantial amount of information generated by these networks, forming relationships with social networks that might be described as symbiotic.

For example, Twitter makes individual tweets and associated meta-information available via a specialised API. According to Twitter Director of Platform, Ryan Sarver[7], as of 2011, there were about 750,000 registered applications accessing the data streams provided by Twitter. Development hotspots include:

Publisher tools: Companies such as SocialFlow help publishers optimize how they use Twitter, leading to increased user engagement and the production of the right tweet at the right time.

Curation: Mass Relevance and Sulia provide services for large media brands to select, display, and stream the most interesting and relevant tweets for a breaking news story, topic or event.

Real-time data signals: Hundreds of companies use real-time Twitter data as an input into ranking, ad targeting, or other aspects of enhancing their own core products. Klout is an example of a company which has taken this to the next level by using Twitter data to generate reputation scores for individuals. Similarly, Gnip syndicates Twitter data for licensing by third parties who want to use our real-time corpus for numerous applications (everything from hedge funds to ranking scores).

Value-added content and vertical experiences: Emerging services like Formspring, Foursquare, Instagram and Quora have built into Twitter by allowing users to share unique and valuable content to their followers, while, in exchange, the services get broader reach, user acquisition, and traffic.

Social CRM, entreprise clients, and brand insights: Companies such as HootSuite, CoTweet, Radian6, Seesmic, and Crimson Hexagon help brands, enterprises, and media companies tap into the zeitgeist about their brands on Twitter, and manage relationships with their consumers using Twitter as a medium for interaction.

According to this classification, TrendMiner is located in the Social CRM or B part of this ecosystem, and therefore well located in an area, where Twitter is actively encouraging development.

2.3 TrendMiner and Social Media

Online media and user-authored content (e.g. weblogs, Facebook and of course, Twitter) are nowadays a major platform for the exchange of information, which is not necessarily only of personal nature. We see that many public personalities, like politicians, also use Twitter for getting messages out to their clientele, and journalists or financial analysts also increasingly rely on alternative means of communication than the "classical" online platforms of email or websites.

The form and the increasing volume of such social and user-authored content has led to challenges of how to access and interpret these strongly multilingual data, in a timely, efficient and affordable manner. In the case of Twitter, this fact gave rise to many of the above mentioned applications. Due to the fact that authors of such

[7] https://dev.twitter.com/blog/changes-coming-to-twitter-api (last access 29.10.2012)

content rarely spend a lot of time composing messages, the level completeness and correctness of communications via Social Media is often low. Messages are short and noisy and their interpretation often require references to other equally short and noisy messages, which might have been posted by the same person or by any other source. Language technology and semantic analysis are confronted with serious data quality issues here. TrendMiner is addressing this type of scientific challenge and aims at delivering innovative, portable, real-time methods for cross-lingual mining and summarisation of large-scale stream media. Summarisation is important since one cannot expect to have tools analysing every messages, but rather to first have a clustering or classification of topics extracted in an efficient and robust way from a large amount of streaming data.

TrendMiner aims at achieving its goals through an inter-disciplinary approach. It combines deep linguistic methods of text processing, to be applied on summaries, knowledge-based reasoning from web science, machine learning, economics and political science. The last two points are related to the use cases, financial decision support (with analysts, traders, regulators, and economists), and political analysis and monitoring (with politicians, economists, and political journalists), that are validating TrendMiner results.

2.4 Classification of Twitter Users

Since TrendMiner is dedicated particularly to the detection of opinions, sentiments and trends developed through social networking and microblogging services offered by Twitter, we take a short look at its users. Users of Twitter (and of Social Media in general) are a heterogeneous group, ranging from occasional to very frequent, from amateur to experts, etc. We present a classification of Social Media users, proposed in the context of "Social Technographics"[8] , displaying the distribution of users in the form of a ladder.

[8] http://blogs.forrester.com/gina_sverdlov/12-01-04-global_social_technographics_update_2011_us_and_eu_mature_emerging_markets_show_lots_of_activity (last access: 29.10.2012)

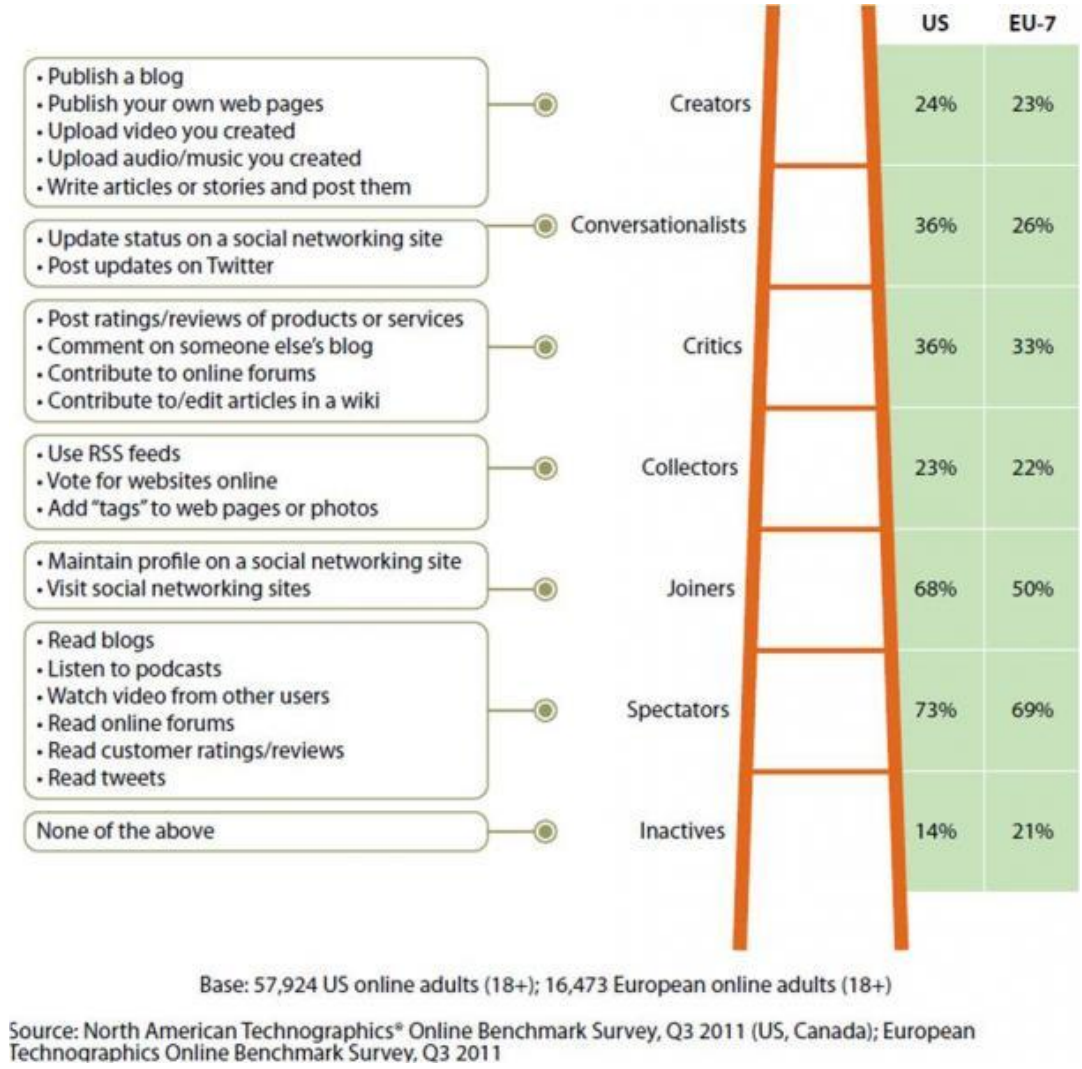

Figure 3: Ladder of activity levels of users of Social Media[9]

In Li and Bernoff (2011), this classification is applied in more details to twitter users:

[9] Illustration taken from Li & Bernoff (2011)

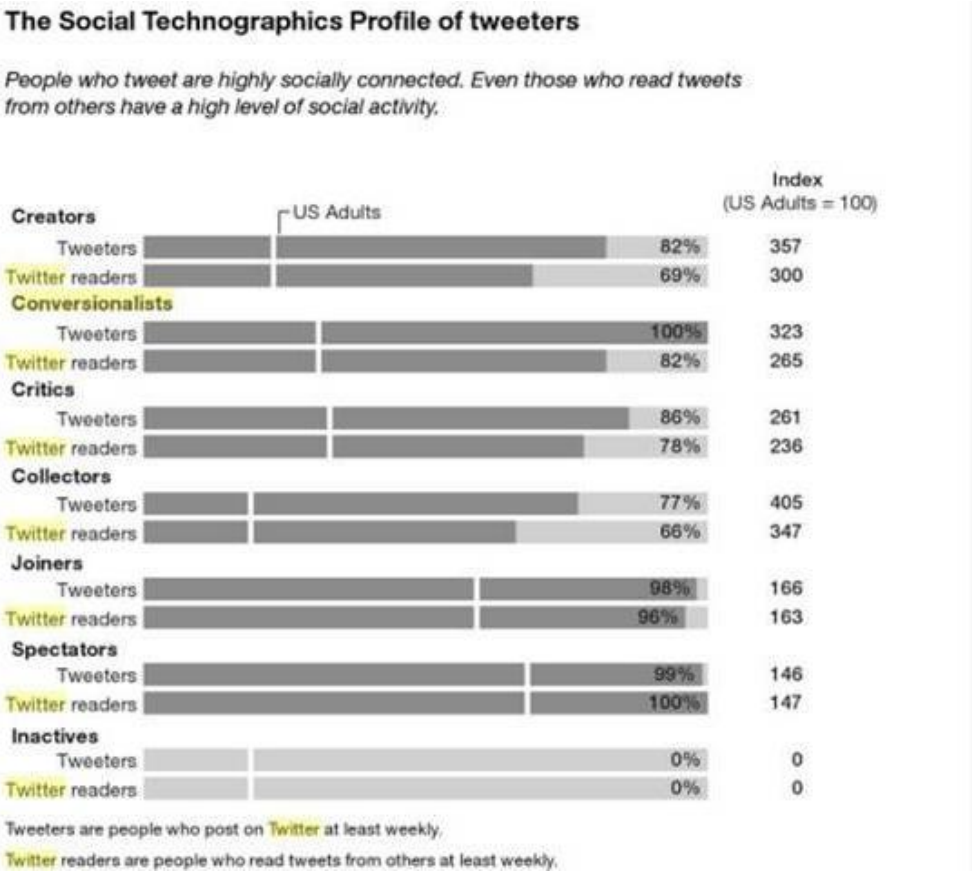

Figure 4: Profile of tweeters[10]

The authors notice that even if only a small part of the society is using Twitter, it turns out that a large proportion of Twitter writers are very influential individuals. This is corroborated by our own observations, e.g. through interviews with party managers and PR professionals who deal with Twitter on a daily basis. Among others, users relevant for Trendminer range from journalists and politicians to engaged citizens, company directors and stock market traders.

The high level of influence that can be found among users of Twitter is one of the reasons, why Twitter is privileged in our selection of data sources for TrendMiner.

[10] Illustration taken from Li & Bernoff (2011)

3 The Market for Services based on TrendMiner

3.1 General Overview

The target market of services based on TrendMiner will consist of investors, traders, analysts, financial institutions, industries, politicians, policy makers, journalists, bloggers and interested citizens. Apart of this, it is clear that social scientists could also benefit from TrendMiner results, since it offers a way to collect substantial amounts of analysed data, giving insights on social interaction in Social Media, political attitudes or the communicative part of financial markets. The market analysis of TrendMiner will consider not only the use of Social Media for content and sentiment analysis, but will investigate the possibility of also covering more conventional online media streams (such as comment sections of online newspapers, etc …).The most relevant benefit of TrendMiner is the ability to develop an accurate detection of opinion and sentiments on a variety of subjects related to the financial and political sphere, also through Social Media, that may be shared by large communities.

3.2 Business Environment

Sentiment analysis can be seen as a market segment of Enterprise Performance Management and Content Analytics. Business Intelligence applications are predominant in this market. Many of these applications are Business Intelligence (BI) and Corporate Performance Management (CPM) tools serving purposes such as CRM, controlling, data warehousing, knowledge management and collaboration. Many of them use techniques also used in TrendMiner, such as text analytics, regression methods, stream reasoning, etc. Such functions are also centrepieces of many Social Media Monitoring applications. Currently, most of the services and consultancy offered in the market for these applications focus on marketing professionals from either companies that are branching out their sales channels into Social Media, or companies that base their business plan on eCommerce from the beginning.

According to Gartner Inc. (Fenn et al 2009), a content analytics application can be a single function, such as keyword extraction. More often, it is a complex function, such as sentiment or trend analysis, fact extraction or reputation analysis. Web 2.0 and business intelligence (BI) is an umbrella of other technologies, it refers to a set of Web 2.0 collaboration and communication technologies such as Really Simple Syndication (RSS), Representational State Transfer (REST), Ajax, blogs and social networking. Web 2.0 and BI provides the ability to synthesize, tag and share information, and deliver analysis more easily, with a greater variety of delivery methods for a greater variety of use cases and deliver a richer, more collaborative experience.

3.3 Market Size

While there is no hard data on the market for tools that provide content analytics or Social Media monitoring could be found for the purpose of this report, it is possible to draw parallels to Business Intelligence and performance management platforms. In this field software revenue reached $10,5 billion in 2010, a 13,4 % increase from

2009 revenue of $9,3 billion, according to Gartner. The four large "stack" vendors (SAP, Oracle, IBM and Microsoft) for general business intelligence platforms continued to consolidate the market, owning 59 percent of the market share. In the BI platform and CPM suite segments, they hold close to two-thirds market share, while in analytic applications, SAS dominates the market. These data are confirmed also by Grimes (Grimes 2011) , who claims that user adoption of Content and Text Analytics grown at a very rapid pace, an estimated 25% in 2010, creating an $835 million market for software tools, business solutions, and vendor supplied support and services. This estimate covers software licenses, service subscriptions, and vendor-provided technical support and professional services. The search-based applications category is worth an estimated $300 million of the $835 million text analytics total value. It includes Web and enterprise search, e-discovery, and business, scientific, and legal information services.

No single solution provider dominates the market for tools that provide content analytics, or Social Media monitoring. Players range from the largest enterprise software vendors to a stream of new entrants, both commercializing research technologies and bringing solutions to new markets. This leaves room for applications based on TrendMiner.

According to Gartner, Content Analytics, in the 2009, was actually positioned at the “post peak” stage of Hype Cycle for Enterprise Information Management and will be a mainstream application within the next 5 to 10 years. This implies that when the TrendMiner project ends, market enlargement will be in process, making commercial introduction feasible and economical and financially sustainable.

4 The TrendMiner Software

4.1 Possible Users of TrendMiner

Three rather distinct groups of potential users of TrendMiner can be anticipated, each with distinct sets of interests and requirements regarding possible use features.

4.1.1 The general Public

This heterogeneous group of users would most likely approach a TrendMiner tool from the same perspective as other, similar tools that are out there on the web[11]. Their main motivation for use would be casual interest. They might take a look at certain topics of interests or explore trends and sentiments visualised by TrendMiner on a daily basis, or just once because they were sent a link by a friend over Twitter and wanted to take a look.

4.1.2 PR or Marketing Professionals

Among this group we might find users such as PR executives, journalists, politicians or party managers etc. Some of its members may be highly active in their professional use of Social Media. This group of users would use a TrendMiner tool as a quick and easy way of exploring trends and hot topics on Twitter or other Social Media because it is professionally relevant for them. The TrendMiner consortium is very interested in this group because they constitute a very important segment of both users and potential customers.

4.1.3 Scientific Researchers

This is a group of users who need a TrendMiner tool to be not only a powerful and reliable way of visualising streaming media, but also to be a scientific research tool, such as social researchers at SORA or financial researchers at EK or HFPL. Their use-mode would require high levels of customisation and comprehensive recording functions.

4.2 Basic Functions

In the course of TrendMiner, a basic Tool will be developed to help validate the results of the technological development. Some basic functions which will be available are[12]:

- Keyword search
- Graphical timeline of topics and sentiment

[11] Examples of such applications are (only to name a few):
https://tool.opiniontracker.net
http://visualization.geblogs.com/visualization/cancerconversation
http://twendz.waggneredstrom.com
http://twittratr.com
http://election.twitter.com
http://twittersentiment.appspot.com/
http://socialmention.com/

[12] More advanced functionalities, like reasoning and queries for a knowledge base are under implementation, but those will probably made accessible only to domain experts.

- Other graphical analysis features, e.g. for number of mentions, level of emotionality, associated topics
- Drill-down to raw data
- Options to filter search results by Time and Date, Places and Languages

4.3 Technological Background:

During the development process of TrendMiner, a number of technologies are involved in the processing and delivering of data and results. Components which are are currently being worked on are described briefly in this section (but more details are available in the Deliverables of WP2, WP3, WP4 and WP5). All components will be integrated in a platform, which will cover all the phases from the (social) stream processing lifecycle: large scale data collection, multilingual information extraction and entity linking, sentiment extraction, trend detection, summarization and visualisation.

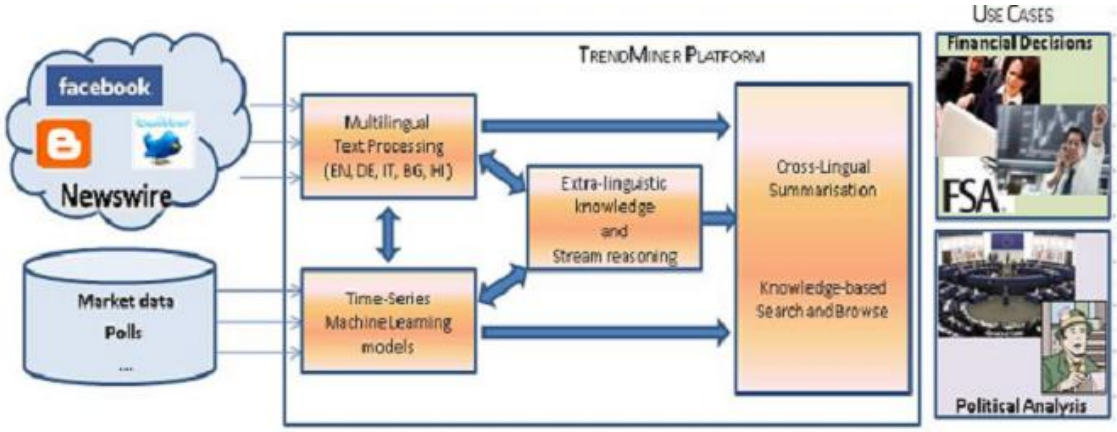

Figure 5: TrendMiner Architecture

4.3.1 Opinion and Sentiment Mining

Detection of sentiments and opinions in the two fields covered by the use cases is a central aspect of the project, and TrendMiner is developing and using novel weakly supervised machine learning algorithms for automatic discovery of correlations between textual data and factual information, supporting the detection of trends. Scalability and affordability are addressed through a cloud-based infrastructure for real-time text mining from stream media.

In the field of Human Language Technology (HLT), opinion and sentiment analysis are playing a steadily growing role. This is partly due to the fact, that the detection of such phenomena in an automated way can show that the natural language processing is getting closer to the task of language understanding, adding the capability to detect high-level contextual semantics that is inherent to language, and not directly dependent from the kind of semantics one can access in knowledge repositories. This increasing interest in opinion, sentiment and/or subjectivity detection in natural language expressions is demonstrated by a high number of conferences, workshops, shared tasks and evaluation campaigns, dealing with opinion and sentiment detection

and analysis, resulting in a huge quantity of available data sets for training and testing systems[13].

One feature making opinion and sentiment analysis highly interesting for HLT is the fact that opinion and sentiment detecting is often applied to Social Media, where many different kinds of textual data so far untouched by HLT projects can be found.

4.3.3 Data Collection and Crawler Service

Working on TrendMiner enables project partner IMR to improve their web-scale crawler, the result of which would be of great interest for several institutions who require real time big data (News, RSS), including Social Media (blogs and forums). IMR would adapt its current crawler to capture real time and social data and its current architecture to use and store this data. The challenge is to do this at scale, to enable real trend mining based on a large number of sources (millions of RSS items, forums and blogs with many comments) while using limited amount of resources and storage to make this a viable service platform.

The adaptation includes developing algorithms for identifying relevant media information sources in multiple languages, prioritization, ranking and filtering of data sources (WP5). These features on top of existing features of spam detection, data cleaning and data de-duplication make IMR's crawler and platform efficient in handling relevant real time and social data and produce high quality repository of data for the consumer to work on. For example concerning blogs and forums, it is important to be able to identify the structure of this kind of website, in order to capture all comments.

4.3.4 Ontology-based Information Extraction

Information Extraction (IE), a form of natural language analysis, is becoming a central technology for bridging the gap between unstructured text and actionable knowledge. Ontology-Based IE (OBIE) is IE which is adapted specifically to the challenge of annotating unstructured content with respect to a formal ontology (a formal conceptualisation). In particular, TrendMiner is addressing the challenge of using large-scale Linked Open Data resources, which contain millions of instances, as well as formal classes and relationships between them.

The automatic semantic annotation of Social Media streams enables the semantic-based summarisation, search, browsing, and visual analytics techniques also developed in TrendMiner as part of WP4. Such knowledge is also needed for building semantic models of the user, their social network, and online behaviour. TrendMiner's LOD-based IE technology is thus relevant in many application contexts, going beyond the use cases pursued in the project, e.g., knowledge management, competitor intelligence, customer relation management, eBusiness, eScience, eHealth, and eGovernment.

4.3.5 Reasoning over Streams

Engaging actively with high-value, high-volume, and dynamic Social Media streams has now become a daily challenge for both organisations and ordinary people. TrendMiner is developing intelligent Social Media summarisation methods that can automate, at least partially, this process. State-of-the-art automatic text summarisation

[13] e.g. International AAAI Conference on Weblogs and Social Media (icwsm.org/) provides data sets for training and testing

algorithms have been developed primarily on news articles and other carefully written, long documents. In contrast, user generated content tends to be very different: often short, strongly grounded in context, temporal, noisy, and full of slang.

To make use of large collections of data, a key requirement is the ability to query the data and sometimes to reason over the data. Though often presented separately, in a very broad sense both querying and reasoning can be thought of as a filtering activity over data and a combination of this filtered data to extract or recombine to produce novel data. The structure of such filter/construction activities are defined in a query language in the querying context or as rules in a reasoning context. Many scientific and commercial activities can produce high throughput streams of structured data. Such activities may include sensor network activity monitoring, Social Media streams, telecom call recording, financial transactions etc. The data produced by these activities is often structured in semantic web data technologies such as RDF. It is often desirable for such data to be queried and reasoned over in real-time, however doing so presents many interesting challenges. Firstly, treating the stream as a standard static Data Base Managment System (DBMS) can prove difficult. One strategy might be to fill a traditional static DBMS with time-separated windows of the stream, though in doing so questions arise as to what happens to the results of a query or entailments of reasoning rules when the stream (by definition) generates more data over time. It is clear that Data Stream Managment Systems (DSMS) represent a separate problem space and over the last decade DSMSes have generated a great deal of interesting research activities. Southampton is developing a framework which leverages the well-studied Rete reasoning algorithm and implements it over a distributable streaming processing framework called Storm. This ReteStorm provides a principled means by which to compile both reasoning rules (e.g. Jena Rules orother rule languages) and queries (e.g. SPARQL, C-SPARQL) as distributed reasoning networks which can be applied to streams. Our implementation supports distributable filters and joins which means our stream reasoning techniques is built to scale with extremely high yield streams, a common problem in many modern streaming data sources. Further, we use a “sliding window” technique wherein joins forget data after certain time period, after some memory limit is reached or some other criteria. This allows ReteStorm to efficiently match queries and produce entailments over a potentially infinite stream.

4.3.6 An Open Source, Social Media Text Processor

With the rise of large online social networks (OSN) there has been an equal rise in the desire to analyse, process and understand activities across these networks. This can be a quite challenging task especially considering the unique issues presented by OSN data including: short messages with little structure; inconsistent spelling, grammar and capitalisation; threaded discussions with many participants across large graphs and so on. Standard text processing tools have been shown to have great difficulty in dealing with such data and therefore many authors have explored algorithms and techniques which purport to handle the unique challenges set by such data. However, these approaches have often been adhoc, implemented in a piecemeal, closed-source fashion usually to address an individual problem and, amongst their other issues; do not address the scale of the data available in modern OSN. To address this gap, Southampton and Sheffield have developed the TrendMiner preprocessing Tool. The tool contains implementations of various components expected in a low level text pre-processing pipeline specifically geared to deal with data from OSN. The techniques provided include tokenisation, language detection, stemming, Pointwise mutual

information and location detection with future plans to provide open implementations of sentiment analysis, POS tagging and Name Entity Recognition. Furthermore the tool provides support for multiple social network formats through support of USMF1, filtering functionality and varying output formats. The tool is engineered in a modular manner, allowing for easy extension and addition of novel techniques. Furthermore, the pipeline of the tool is specifically designed to work either as a library, a single machine command line tool or even to analyse massive datasets with a MapReduce enabled implementation of the tool. DFKI is developing tools for the automatic extraction of polarity lexicons out of newswires, and grammars for computing the polarity of phrasal and sentential units.Addtional to this, DFKI is implementing domain ontologies, a biography ontology and an opinion/sentiment ontology. The combination of those ontologies allows to store consolidated information about opinionated entities, resulting from either information extraction, stream reasoning or summarization.

5 Existing Content Analytics Providers

5.1 General Introduction

Content analytics and Social Media monitoring are growing branches of software services, there is a large variety of providers. All of them have in common, that they tap the huge mass of content and data available on the web through sources like Social Media, online newspapers, stock exchanges, online forums and the like. Another parallel is their general purpose: to aggregate, structure and visualise the constant data flows. Using different ways of analysis and summarisation, including natural language technologies, they produce information on a scale that a human user can manage. They differ however, in the exact use to which they are put, in the details of the sources they access and the analytical options they offer.

We now present selected examples of important providers of content analytics and their products.

5.1.1 ExactTarget

"SocialEngage (formerly CoTweet) is a Social Media publishing and engagement platform built for the unique needs of businesses of all sizes. SocialEngage helps marketers, agency professionals, and customer service teams publish content and manage their day-to-day Social Media conversations on multiple social networks from a centralized, easy-to-use dashboard." SocialEngage was developed by ExactTarget, a global company which is headquartered in Indianapolis, Indiana (USA). It was founded in 2000 and now has more than 1,200 employees spanning four continents. The company is also a recipient of awards and accolades from third-party organizations.

ExactTarget provides solutions for Email-, Mobile- and Social Media Marketing and has an Online Technical Library for developers. For Social Media Marketing purposes it analyses customer interactions on Twitter and Facebook.

ExactTarget sets a high value on the security of its application, infrastructure, and its clients' data, as part of its initiative called "Protected by ExactTarget"

While it accesses both Facebook and Twitter, the core business of ExactTarget is rather supporting interaction with users of these services (e.g. customers and employees), or collecting and managing data about them than providing an in-depth analysis of these data as TrendMiner does.

5.1.2 HootSuite

HootSuite is a global, privately held company with more than 150 employees and which is headquartered in Vancouver, Canada. It got funding from Blumberg Capital, Hearst Interactive Media, Millennium Technology Ventures and investors Social Concepts and Geoff Entress. The company also received a number of awards and accolades from the Mashable Awards, Shorty Awards and Digi Awards.

HootSuite provides solutions for managing all of one's social networks (Twitter, Facebook, LinkedIn, Google+, Foursquare, MySpace, Wordpress, Mixi, and with App Directory: Tumblr, YouTube, Flickr, as well as marketing focused tools such as MailChimp, SocialFlow, InboxQ, Constant Contact, and many more), custom analytics (this function is the nearest equivalent to TrendMiner due to its analysing

functions), team management, communication within organisations, and message management in social networks.

5.1.3 datenwerk innovationsagentur GmbH

Opinion Tracker is a service provided by datenwerk innovationsagentur GmbH which is a private limited company and headquartered in Vienna, Austria. The commercial object of the company is consultation, conception, development, and sale of computer-aided methods for interaction with clients, the construction of communities and provision of services in the fields of Internet and new media. The main target communities are businesses, politicians, institutions and organizations.

Opinion Tracker is a proprietary product which covers German and English online news and Social Media. Both single and enterprise level subscriptions are available for a monthly fee. The software enables an interactive visualisation (including a timeline), a search function, track sharing, ad hoc analysis (most important people, topics, and organisations), and a daily Email status update. It is similar to TrendMiner insofar it has many analysis features that are desirable for TrendMiner. On the other hand, it is focused on the german speaking countries Austria, Germany and Switzerland and offers only English as an additional language.

5.1.4 RavenPack

"RavenPack is the leading provider of real-time news analysis services. The company's clients include some of the best performing quantitative and algorithmic trading firms in the world" [14] .RavenPack's products are aimed at analysing news through linguistic methodologies. RavenPack's analysis is based on hundreds of thousands of stories per day coming from several sources in different formats in milliseconds, RavenPack's news analysis is also able to identify short-term trends.

The RavenPack's product for Content analysis is named News Analytics that automatically monitors and analyses information on over 100 countries and governments, more than 140,000 key geographical locations and 30,000 companies. RavenPack Analysis also collect more than 1,000 types of events related to corporate actions, terrorist threats and natural disasters. The news related to events are categorised and the sentiment analysis is based on different metadata divided in terms of place, organisation, company, currency or commodity. RavenPack News Analysis service is in partnership with Dowjones.

5.1.5 Dow Jones

"Dow Jones & Company explains the world and the world of business. With authoritative journalism and smart technology, Dow Jones provides a window on events, clarify issues, inspire new thinking and give readers and business customers the insight they need to make informed decisions" [15].

Dow Jones & Company is a big company of journalism and smart technology. Dow Jones owns several newspapers, newswires, websites, apps, newsletters, magazines, proprietary databases, conferences and more. Dow Jones' premier brands include: The Wall Street Journal, Dow Jones Newswires, Factiva, Barron's, MarketWatch, SmartMoney and All Things D. The Dow Jones Local Media Group publishes community newspapers, websites and other products in six U.S. states.

[14] http://www.ravenpack.com/aboutus/index.htm
[15] http://www.dowjones.com/about.asp?link=djc-topnav

Dow Jones' products are mainly focused on testing short- and longer-term algorithmic trading models by analysing news and data archives.

Dow Jones News & Archives For Algorithmic applications: this is one of the two Dow Jones' products for news analysis and it is aimed at mining news and data from Dow Jones' 30 years archive through algorithmic trading models[16]. This product covers also the sentiment analysis.

Dow Jones News Analytics: this product analyses real-time and historical news and data for sentiment through trading models and it is able to analyse the relevance, volume, novelty and other market signals via a range of technology options.As mentioned above, the Dow Jones applications are apparently based on technology developed by RavenPack, and are therefore most likely very similar in makeup and capabilities. Both Providers seem to be focused on analysing classical news streams rather than data streams from Social Media, making them significantly different from TrendMiner in terms of data source.

5.1.6 Thomson Reuters

"Thomson Reuters is the world's leading source of intelligent information for businesses and professionals. It combines industry expertise with innovative technology to deliver critical information to leading decision makers in the financial, legal, tax and accounting, healthcare, science and media markets, powered by the world's most trusted news organization" [17].The Sentiment Analysis product by Thomson Reuters is named News Analytics (TRNA) for Internet News and Social Media and is a tool aimed at analysing millions of public and premium sources of internet content, tag and filter. The tool then turns the data into actionable analytics in real time to support trading, investment and risk management decisions. The TRNA engine is mainly deployed by trading firms to analyse Thomson Reuters' News and a host of professional news wire services. TRNA for Internet News and Social Media product aggregates content from more than four million Social Media channels and 50,000 Internet news sites. The TRNA engine is also able to analyse not only sentiment, but also relevance and novelty. Thomson Reuters and its News Analytics product seems to be the most similar provider to the TrendMiner objectives. It analyses, in real-time, more than 50000 aggregated news and more of 4 million Social Media channels. The main difference is in the business model behind the Thomson Reuters services and the aim of the TrendMiner project. One difference lies on the outputs of these services: Thomson Reuter generates an output of quantifiable data points across a number of dimensions, but it does not offer directly forecast on the market through proprietary tools; TrendMiner aims to provide sentiment and, at the same time, a forecast on the markets through a financial model ad hoc developed. Another difference regards the provision of the main functionalities of TrendMiner for free to a public audience, and while Thomson Reuters is an expensive service and its completely provided in order to make profits.

[16] An Algorithmic Trading is "a system that utilizes very advanced mathematical models for making transaction decisions in the financial markets (Investopedia: http://www.investopedia.com/terms/a/algorithmictrading.asp#axzz2AgL0Tkbg) (last access 29.10.2012).

[17] http://thomsonreuters.com/about/

5.1.7 FINIF

"FINIF specializes in financial sentiment analysis which allows investors to derive unique insights from Social Media. Financial sentiment monitoring is an emerging technology and our algorithms make sense of the thicket of SEC[18] filings, myriad news articles, and thousands of tweets pouring in every minute using textual analysis to gauge investor sentiment"[19]. FINIF examines various sources of real time data (e.g., SEC filings, news headlines, and tweets) and use textual analysis to measure the sentiment of the text or flag certain sensitive phrases sensitive phrases in order to measure the information environment for a firm. Financial sentiment monitoring is an emerging technology and our algorithms make sense of the thicket of SEC filings, myriad news articles, and thousands of tweets pouring in every minute using textual analysis to gauge investor sentiment.

FINIF financial sentiment analysis scan for new 10-K and 10-Q20 filings and then apply the sentiment algorithms to create investment's reports. The product evaluates the number of positive and negative news of a specific company compared to the stock price. The FINIF product is mainly based on the analysis of Tweets and information flows of a specific company. FINIF product creates also a sentiment score by aggregating the sentiment measures across a sample of the most recent Tweets. While this product is rather similar to TrendMiner in terms of analysis features and data sources, it lacks multilingual capabilities and is not suited to provide outputs that would facilitate forecasting based on available data.

[18] Security and Exchange Commission

[19] http://www.finif.com/about/

[20] "Form 10-Q, (also known as a 10-Q or 10Q) is a quarterly report mandated by the United States federal Securities and Exchange Commission, to be filed by publicly traded corporations." (Wikipedia - Securities and Exchange Commission Form 10-Q) and " A Form 10-K is an annual report required by the U.S. Securities and Exchange Commission (SEC), that gives a comprehensive summary of a public company's performance." (Wikipedia - Form 10-K).

6 Commercial and scientific Opportunities for TrendMiner

6.1 General Overview

In this section, every partner of TrendMiner offers their view on possible commercial and scientific opportunities given by their contribution to both the R&D work and the use cases. Research partners also compare TrendMiner goals with existing systems.

6.2 USFD

Open-Source, Multilingual Ontology-based Information Extraction

There are a number of commercial online entity recognition services which annotate documents with entities and assign Linked Data URIs to them, similar to the TrendMiner techniques from WP2. Here we list some of the most popular such services:

AlchemyAPI[21] is a commercial service, which offers a limited, closed set of text annotation pipelines, including named entity recognition and disambiguation.

Zemanta[22] is an online semantic annotation tool, originally developed for blog and email content to help users insert tags and links through recommendations. It is then for the user to decide which of the tags should apply and which in-text link targets they wish to add to their blog post.

Open Calais[23] is another commercial web service for semantic annotation. The target entities are mostly locations, companies, people, addresses, contact numbers, products, movies, etc. The events and facts extracted are those involving the above entities, e.g., acquisition, alliance, company competitor. The entity annotations include URIs, which allow access via HTTP to obtain further information on that entity via Linked Data. Currently OpenCalais links to eight Linked Data sets, including its own knowledge base, DBpedia, Wikipedia, IMDB and Shopping.com.

The main limitation of these tools comes from their closed nature, i.e., users send documents to be annotated by the web service and receive results back, but they do not have the means to give the service a different LOD resource to annotate with. Secondly, they have limited support for languages other than English, which is an even bigger shortcoming and, conversely, a key strength of the TrendMiner technology.

6.3 SOTON

Reasoning over Streams

Engaging actively with high-value, high-volume, and dynamic Social Media streams has now become a daily challenge for both organisations and ordinary people. TrendMiner is developing intelligent Social Media summarisation methods that can automate, at least partially, this process. State-of-the-art automatic text summarisation algorithms have been developed primarily on news articles and other carefully

[21] http://www.alchemyapi.com
[22] http://www.zemanta.com
[23] http://www.opencalais.com/

written, long documents. In contrast, user generated content tends to be very different: often short, strongly grounded in context, temporal, noisy, and full of slang.

Summarising sentiment in tweets and product reviews is one of the most frequently addressed Social Media summarisation tasks. As mentioned earlier, there are a number of commercial online services which offer sentiment-based quantitative summaries, mostly as percentage positive vs negative tweets towards a given person, company, or brands.

These services classify the tweets as positive or negative. However, this is not with respect to the specific entity that is being searched for, but captures the sentiment of the tweet as a whole. For instance, after Witney Houston's death, TwitterSentiment reported 83% negative sentiment towards her. In reality, it was people being sad about her passing away.

Going beyond tweet sentiment summarisation, researchers have worked on text-based product review summaries, and in particular, aspect-based summarisation and ultra-concise pros and cons summaries. However, these methods are currently in early development stage.

Therefore, the WP4 Social Media summarisation technology has the potential to advance significantly what is currently possible and available on the market. In the first instance, the techniques are developed and tested on the political and financial use cases, but they have a much wider relevance, including brand and reputation management, customer relationship management, competitor intelligence, and other applications, where large amounts of Social Media streams need to be analysed and acted upon.

As discussed, various activities may produce streams of structured data. Our system provides a principled way to process any of these streams. Administrators of such data streams may look at solutions like ReteStorm to reason over streams directly where previously they may have saved chunks of their data and attempted to marry the query results and reasoning outcomes across these chunks. Concretely, consider the Social Media streams handled within TrendMiner. Streams such as twitter produce 200K tweets per second, but individual subjects of interests have been recorded to produce over 110K tweets per minute on their own when significant events such as the US elections are taking place. Though real-time collection of such volumes of data is a challenging task of itself, real time processing can present an even greater challenge. One approach to addressing such problems are to scale processing in a distributed way, ReteStorm could be used to address such styles of processing.

6.4 IMR

Social and Real Time Media become an essential source of information. It would not be relevant anymore to only focus on traditional media – TV, newspapers, websites and etc. Hence the focus must be onleveraging special publication characteristics of this type of content, specifically RSS and APIs. RSS can be detected based on large crawls (that Internet Memory current performs) and distilled by topics and relevance to enable real-time probing for new content, and rapid capture and processing where required. So as a sourcing provider IMR needs to adapt its technologies and processes to social and real time media sources.

The adaptation includes developing algorithms for identifying relevant media information sources in multiple languages, prioritization, ranking and filtering of data sources (WP5). These features on top of existing features of spam detection, data cleaning, data de-duplication and archiving would make IMR's crawler and mignify platform efficient in handling relevant real time and social data feeds and produce high quality repository of data for the consumer to work on. For example, and concerning blogs and forums, it is important to be able to identify the structure of this kind of website, in order to capture all comments.

Any research centre which work on social and real time data (for example: research on Social Media or research on consumer behaviour) would be interested in working on such real time and clean data set.

IMR is currently involved in couple of research projects namely DOPA (funded by the European Commission, N° 296448) and AnnoMarket (funded by the European Commission N° 296322). The technological advancements in TrendMiner would be of interest for the other projects and vice-versa.

6.5 DFKI

In the field of Human Language Technology, opinion and sentiment analysis are playing a steadily growing role. Similar to USFD, the language technology lab is interested is working on Ontology-based Information Extraction. With a focus on linguistic aspects. Extracting automatically polarity vocabularies from streaming media is a challenge, which is leading to a dynamic and contextual view on lexicons to be used in HLT applications. DFKI is also interested in providing means for semi-automatic acquisition of (shallow) ontologies from domain documents and to link those to ontologically organized opinion/sentiment features. An outcome of the work of DFKI in the project is the TrendMiner Set of Integrated Ontologies, relating ontologies in the field of finance, politics, biography, all equipped with temporal and opinion features. The fact that we invest work in the development and integration of ontologies is due to the fact that the main scientific credit of TrendMiner lies in the combination of high-level semantic representation and very robust distributed statistical approaches to analysis of natural language expressions, available in various languages, formats and quality.

But apart from the scientific community, that there are a lot of institutions and industries interested in the outcome of projects like TrendMiner, for receiving an as accurate as possible detection of opinion and sentiments as shared by an as large as possible community. This can lead to modify the direction taken by politicians and people involved in development and marketing of products.

While TrendMiner is dealing with multilingual input, it does not provide for translation. This can be achieved only with a close collaboration with projects, groups and companies dealing with hybrid translation systems in the future. The same holds true for linked data (LD). It would be nice to have results of TrendMiner combined with the type of semantic distribution offered in the linked data framework, and adding opinion or sentiment feature to facts encoded in the LD. Here again, this can be achieved only in the context of closer cooperation with relevant players in the LD field.

We already mentioned politicians (and any government agencies) and people defining strategies for the development and distribution of (new) products. Apart of this, it is

clear that social scientists could benefit substantially from TrendMiner results, offering substantial amounts of analysed data that can be used for validating models in use in social sciences.

6.6 EK

In the last few years, the financial markets have seen more and more uncertainty and volatility of prices due to the general economic conditions. The first consequence is faster changes of the expectations in the short periods. Recently, it's been possible to observe how policy announcements changes have caused wide variations and trend reversals in a very short time, like on the bond markets and their spreads. This increasing complexity has as counterpart the most important role which information plays. In addition to the traditional information flow coming from official channels (e.g. balance sheets, central banking and stock exchanges reports, official news), specialised web sites, analyst blogs, tweets, rumours, opinions and their worldwide links (i.e. the information from Social Media) play a more and more important role in turning the "average" opinion towards herding or contrarian collective behaviour, supporting bubbles or crashes. The combination of the increase uncertainty and greater attention to all sorts of news or rumours, has also led to investors' behaviour often moved by emotional states, away from rationality postulated in many of the financial models used in the financial forecasts. While investors' questions are always the same (What are the impacts of an economy-related political announcement? Is the governmental plan reliable? How will citizens modify their economic behaviour? Will they have positive expectations about the future of the economy? Moreover, is the industrial plan of a public company perceived as good enough? Will the profits be higher? Will financial investors buy or sell the shares? What will be the future path of economic and financial values?), the answers have to recognise both the more influent role and impact on the expectations of news and of the collective behaviour, often irrational.

The strategy of commercialisation and the ability to capture scientific opportunities of EK is based on the previous premises.

From the scientific opportunities point of view, EK would improve its market model based on an agent-based approach. At the moment, this (artificial) market model is able to develop exchanges between agents returning prices time series that have the same properties of the real stock markets (stylized facts). The first improvement is to further develop the model considering the flow of real information (prices and volumes) and, especially, to modify the expectations module in order to use and to extract meanings (i.e. correlations and forecasts from the media streaming content analytics) from the sentiment analysis (i.e. the sentiment index) developed through the TrendMiner project. The goal is to publish this model in international journals with high impact factors and to provide to the research community in finance a platform able to forecast and to analyse the correlation between news, sentiment, financial markets and prices. Academic researchers and financial authorities, for example, might be interested in analysing the sentiment time series to look for correlation and casualties between news and particular movements on the financial markets (e.g. markets distortions).

Figure 6: Current interface of the Eurokleis trading system software

From the commercialisation point of view, the potential users could be: analysts, traders, financial investors, citizens and consumers, financial authorities, journalists, CFOs, advisors, stock brokers. The supply of EK should aim to satisfy the following needs:

- having a comprehensive idea about the sentiment on the Social Media towards financial instruments / listed companies (improving the completeness of the information available)
- receiving indication about the potential prices movement path through the correlation with the sentiment index, in order to positively modify their economic behaviour, to detect and deliver only actionable trends critical to their specific environment
- accessing and interpreting strongly multilingual financial data, in a timely, efficient, and affordable manner
- allowing all brokers to afford in-house research
- taking advantage of a methodology for predicting the appropriate time to buy or sell a stock using technical indicators, such as: volumes, highest and lowest prices per trading period, charts, moving average and, especially, the sentiment index developed by TrendMiner.

The commercial target can be clustered in three main groups:

1. The general public: this group includes all persons interested in the sentiment summarisation, from Social Media, regarding a listed company. They could be not really interested in the (potential) subsequent movement of stock price, but they should be really interested to have a clear and easy idea about the "social" sentiment is. This group includes also companies' management, journalists, and consumers.
2. Private/Retail Investors: this group includes all persons interested in the sentiment index as input for their financial decisions. They are potentially interested in using the sentiment index like an alert indicator to signal a financial momentum basing on its correlation with trends and/or price

movements. Due to their small capital, maybe they could be interested in taking position on the midterm of the financial markets and, consequently, they could be interested in the sentiment analysis calculated on daily basis.

3. Professionals: similar to the previous group, the professionals are interested in having a sentiment index computed in real time to be implemented in an automatic algorithm or to use sophisticated models, sentiment based, to forecast the price movements of particular stocks.

6.7 SORA

Information about Opinion Leaders

From current research (Maireder et al 2012, Maireder 2011) as well as from interviews we have conducted with Party Managers and PR professionals, we know that Austrian Twitter users represent a powerful mix of Journalists, Bloggers, Politicians and other political activists. The user community on Twitter may not be representative for the whole of Austrian society but many opinion leaders and multipliers can be found among them. This mix is likely present among most, if not all, language communities of Twitter users.

Users interact closely, exchanging information and opinions, conducting a public negotiation of news spin. As one party manager interviewed by SORA puts it: "[..] many journalists in a way anticipate their comments for tomorrow on Twitter, or rather anticipate the spin they will give their articles".

The unique and influential user base that is present onTwitter, combined with TrendMiner´s fast analysis capabilities will grant unprecedented and immediate insights into the day-to-day formation of public and media opinions.

Fast Analysis and Reaction

Professionals working in the field of political PR are often faced with the need to identify discussions and public opinions or moods very quickly. They need to react in a focused and efficient manner. Currently existing tools that address Social Media monitoring automated content and sentiment analyses in real-time are not specialised to address political content but tend to be geared towards brands or financial services.

In TrendMiner the political use case is one of two major development focuses and thus closes this gap. It offers functionalities to political PR professionals that were hitherto only available for classical Marketing and eCommerce purposes.

However, the nature of the user base present in Social Media like Twitter also limits the representativeness of the topics and associated opinions that can be found there. Currently it is not possible to generalise opinions found on Twitter to a whole electorate, because the set of people who use Twitter are neither representative for the electorate nor drawn at random

Speeding up Media Content Analysis

The current best practices in the domain of political analysis consist primarily of either interfacing with news and blog articles only, or interfacing with limited aspects of the realtime web in one language (such as accessing Twitter on one hashtag). Further, processing is not done in real-time and many interfaces to the data are

simplistic (basic positive/negative sentiment) rather than capturing many variables. The NLP and stream reasoning software functions of TrendMiner are intended to speed up this process and can be of substantial assistance to MCA in the political field. TrendMiner also addresses the issue of sentiment and extends this from traditional media into the realm of online Social Media. TrendMiner also supports the analyst with a variety of functions such as timelines, geographical breakdown and content summaries.

Supporting comparative Political Analysis

By transcending language barriers, TrendMiner offers a new approach on comparative research in political analysis. It will be possible to analyse and compare discourses across language barriers, facilitating the analysis of pan-European policy, political communication and campaigns. Public officials on the national and the European level are also presented with an additional means to gauge political trends and sentiment across borders and languages.

6.8 HFPL

A core part of a financial analyst's daily work is monitoring different media sources, such as Bloomberg, TV news channels, newspapers, and websites, in order to detect potential stock volatility and price issues and advise clients accordingly. At HFPL, like many other similar companies, this is a manual process where human analysts spend a lot of their time monitoring media. The heavy increase of available only information as made this harder, and more work intensive.

HFPL therefore stands to gain significantly from new, intelligent stream media mining and summarisation tools to be developed in TrendMiner. Their real time performance and ability to cope with Social Media streams, especially Twitter should translate into significant labour and cost benefits for HFPL, who will also be able to offer an improved, more responsive service to their clients. Any realised efficiency gains will also allow HFPL to widen their customer base.

For HFPL, Multilingualism is among the key benefits here. as it allows analysts to tackle more efficiently the wealth of business web sites which they currently monitor both in Hindi and English (e.g. bhaskar.com). Of particular interest is the ability to process also the comments left by users on the web site, some of which are in English and some in Hindi.

The real-time results produced by the tools developed in TrendMiner will not only save an analyst's time, but also help him or her, to deliver the a sound analysis on time. Currently, we know of no tools that perform real time news analysis and are able to cross over between English and Hindi. HFPL, we plan to use these tools to increase our clients and help our clients take right decisions and get the best real-time analysis.

We also plan to collaborate with other stock brokers in India to provide their customers with the analysis service. We also provide a portfolio management service and we plan to use TrendMiner tools to enhance it.

6.9 ONTOTEXT

Ontotext is a provider of core semantic technology, focusing on performance, scale, and compliance with open standards. It is the developer of OWLIM , the most scalable semantic database. Another outstanding product of Ontotext is KIM – a semantic annotation and search platform. Ontotext will extend its expertise to offering a platform for large real-time stream media collections, their analysis, and indexing. The company wants to strenghten its expertise in cloud computing for semantic analysis and ontology reasoning and linguistic component, with contribution on multi-faceted search of large document collections.

7 Exploitation Goals

7.1 General Overview

In the former sections we described the various challenges and opportunities that the partners of the TrendMiner are confronted with in the project. In this section we summarize the particular exploitation plans of the members of TrendMiner.

7.2 USFD

USFD has considerable experience and expertise in the development and promotion of open-source software in language technologies and related fields, notably the GATE software family (http://gate.ac.uk). USFD provides training and consultancy services for the use, development, and application of its software, including the multilingual information extraction and Social Media summarisation algorithms developed in TrendMiner.

The results achieved and software developed in TrendMiner will be promulgated to the other partners and spin-offs; furthermore, take-up by academic and other audiences, as well as participation in the development and implementation of standards, will yield further exploitation opportunities. The results will also maintain USFD's reputation in language technologies and machine learning for NLP and thus its ability to participate in new projects and further training and consultancy.

In more detail, since June 2012 the USFD team has been working together with the British Library and HR Wallingford on the customisation and use of the English ontology-based information extraction tools, delivered in D2.1.1, in the context of environmental science. These tools are used together with DBpedia, GeoNames, GEMET, and the Ordnance Survey Hydrology ontology, to enrich semantically the metadata and fulltext articles indexed in British Library's Envia information discovery and access tool[24].

USFD has also had preliminary consultancy discussions with the British Red Cross, who are similarly interested in having their internal documents on disaster relief and flooding indexed with relevant semantic terms and entities.

We expect that as the TrendMiner language technologies mature in years 2 and 3, so will USFD's training and consulting activities.

7.3 SOTON

ReteStorm

In the short to medium term Southampton plans an initial paper based on the work on ReteStorm to be published in a high impact conference (WWW'13). In this paper we hope to outline the approach, present initial benchmarks and results as well as present an initial release of the codebase and tool. Beyond this, in the long term we plan to release ReteStorm as Open Source Software (OSS), providing value to future semantic web projects both in and outside of the EU

[24] http://www.bl.uk/reshelp/experthelp/science/eventsandprojects/enviatbl/index.html

Open Source Social Media Text Processor

In the short to medium term Southampton and Sheffield plan to implement and evaluate various additions to the tool, focusing on sentiment extraction as our next primary concern. Beyond this, in the long term we plan to further improve and evangelise the TrendMiner tool as Open Source Software (OSS), providing value to future OSN research both in and outside of the EU.

7.4 IMR

IMR's commercial platform, mignify, has a lot to gain from this project. With the improvement in the crawler to adapt to the TrendMiner requirement, the overall platform would be able to improve its quality of service on all kind of data and feeds (traditional website, real time and Social Media) and reach out to a broader range of customers. Several contacts of IMR (technology providers, market analysis companies…) have expressed their interest in an industrial offer of such sources.

Thanks to TrendMiner project, mignify platform will be able to service Social Media stream requests and could extend its role in the value chain as it can work in multiple languages and filter and rank the sources. This would be of interest for marketing, PR and News agencies monitoring specific entities and mignify could source them real time focused information.

Also, the mignify platform, with its expertise on financial domain from the DOPA project (funded by the European Commission, N° 296448), could integrate the adaptation being built for TrendMiner project to obtain real time financial news from focused source which would interest the financial agencies.

Also, mignify could foresee, partnering with members of the TrendMiner project, who can plug in their technology on the platform. This would result in higher processed or value added data for the consumer.

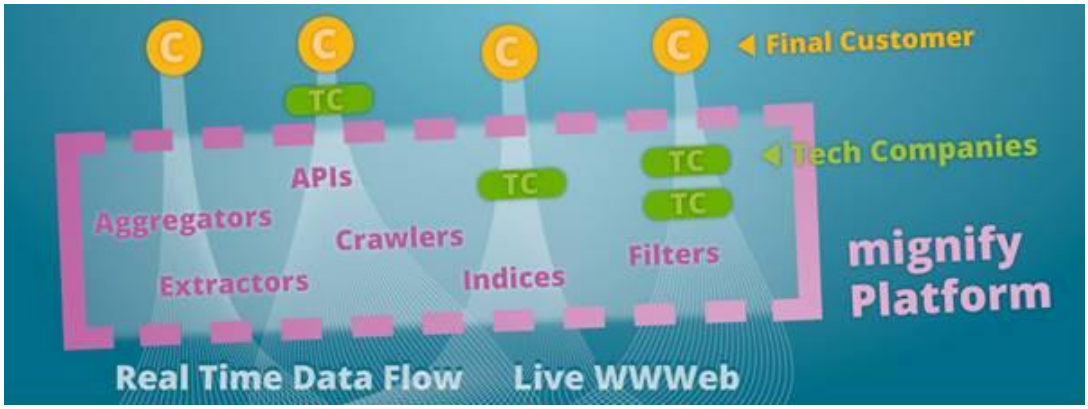

Figure 7: Functionality and Positioning of Mignify platform in the value chain

TrendMiner as a whole has a potential of being a new service with every partner playing a part in building the product.

7.5 DFKI

As a private but non-profit research institute, DFKI has no direct commercial exploitation plans. But indirectly, DFKI is promoting results of R&D projects via the creation of spin-off companies.

It is intended to establish closer technology transfer agreements with the DFKI spin-offs that are potentially interested in the outcome of TrendMiner. A good candidate for this is the company "Attensity"), which emerged from a former spin-off, Xtramind. This kind of transfer agreement is not restricted to spin-offs.

Technology transfer and future research directly funded from industry is the main focus for our group in order to establish a longer term development in the fields covered by TrendMiner. We are in contact with 3 different companies, which are offering opinion or sentiment analysis, but this on the basis of methods not using linguistic or semantic features. In the course of the project, we will compare results of TrendMiner with the results those companies are offering in the demonstrators accessible on their web page. We will be careful here in not entering in conflict with strategies to be implemented by the use case partners involved in TrendMiner.

As the scientific level, DFKI already established contacts and exchanges of data and results with other groups and initiatives, like the IGGSA (Interest Group on German Sentiment Analysis) society. Here we are aiming at generating relevant data sets for supporting advances in the state of the art in NLP-based opinion and sentiment analysis. This is pursued via the generation of lexicons and annotated data sets that can be freely used by other research groups.

Last but not least, we are in contact with two academic institutions dealing with the domains of the two use cases of TrendMiner, and a deeper exchange of methodologies and data sets has already started. An extension to other academic institutions will ensure a longer term research goal.

7.6 EK

The outputs of the TrendMiner project will be exploited by EK to develop several tools and services.

The first tool to be provided to the general public (i.e. international financial community) is an online tool targeted to the categories identified in the previous paragraph, according to the type of customers. It will be a basic account where basic information about sentiment and impact on the price and/or volumes (up/down and the probable future price) is provided for free. Premium profiles, for professionals, will be based on tools from more sophisticated models. The revenues will be generated by advertisement (for free accounts) and/or subscription fee (premium profiles).

A second tool will be delivered as a mobile "app". The goal is to optimize the information flow delivery to the costumers of the online tool, through for example, alert systems. Other (free or premium) apps will be developed for educational, recreational and professional purposes (e.g. e-learning "apps").

Since EK is also a consulting enterprise, on the basis of the tools and models developed by the TrendMiner project, a great importance will have the consultancy

towards banks and financial professionals to set up forecasting models, investing strategies, portfolio optimisations and risk management strategies.

7.7 SORA

An important aspect of SORA´s work in general, as well as a basis of its success, is the provision of consultancy and knowledge based on SORA´s experience and know-how in research methods, as well as the fields of electoral behaviour and political culture.

SORA´s primary customer base are professionals who work in the political sphere, civil society and in public institutions. To support the development of the TrendMiner software, as well as to prepare a more concrete strategy, SORA has approached interviewed representatives of its existing and potential clientele to investigate the way they are using Twitter and other Social Media now in their work, as well as their needs and interests concerning tools like TrendMiner.

Results of these interviews show, that many recognise the need for a tool to analyse and structure the continuous stream of data that is produced by Social Media like twitter. Great emphasis is also placed on the need for a tool that returns prompt results and analyses that reveal the connections between topics, sentiment and opinion holders.

Based on the TrendMiner tool, SORA could offer subscription based, personalised reports on topics of interest on Social Media. Another possible product would be the ex-post evaluation of election campaigns and other activities that are strongly linked to public opinion, based on retrospective analysis of TrendMiner data. A strong analyser of streaming media content can greatly support SORA´s traditional media content analysis activities. TrendMiner also permits a scientific study of the dynamics of the political news cycle on twitter and its relationship with public opinion and classical newsmedia. Finally, TrendMiner may aid the forecasting of election results on the basis that it provides an overview of the landscape of opinions and moods among users of Social Media (who include opinion leaders such as journalists).

7.8 HFPL

HFPL assists their clients by providing analysis of the market on day-to-day basis. At present, communication between the HFPL's market analysts and its clients takes place over phone where the clients are interested in views of the analysts for certain stories released on different media channels such as news on TV, newspapers and internet. Even though people do participate in talks on social networking websites that affect or drive mood in stock markets, it is very unclear at this stage, whether the clients would trust the analysis that takes into account sentiments derived from such sources. Therefore, it was our main goal to educate our clients and make sure that they understand the purpose of the research taking place as part of the TrendMiner project. The Investor Awareness event, held in April 2012, was used as the platform to announce HFPL's involvement in the TrendMiner project. Even though, the clients who had attended the event have applauded the participation of HFPL in TrendMiner and shown great interest in the future outcomes of the project, it is bit difficult to predict how well the technology will be welcomed when it is actually put into the practice. This is true, at least, until the results are evaluated to see whether the

semantic technologies are able to capture sentiment of people and predict its effect on the market.

The investor awareness program was also used as an opportunity for face-to-face communication with the clients to understand different formats in which the analysis should be presented and the preference of different mediums though which the analysis should be delivered. Based on what was discussed at the event and also based on the analysis of the survey undertaken at the start of the project, the company has set the following goals for the coming years. All the plans are subject to positive evaluation of the technology used as part of the TrendMiner project.

Encouraging Investors to invest more

From the past experience and the analysis of the survey undertaken at the start of the TrendMiner project, it would not be unfair to say that the customers who invest higher amounts in the stock market are better risk takers than those who invest less. We, at HFPL, believe that the clients who take low risk at the moment can be encouraged to take higher risks only if they are properly educated about the risks involved and presented with the facts behind the analysis. Thus, positive evaluation of the TrendMiner technology would be a key factor to boost clients' trust in the technology. In order to encourage investors to invest more, the company is planning to introduce different packages of analysis with various features and different modes of delivery.

Attracting Sub-brokers

HFPL has a number of sub-brokers who pay a part of their own commission to the company for the services they rent from it. The company is planning to offer analysis packages as added incentives to their existing sub-brokers and to attract more sub-brokers. One such example of an added incentive would be the terminal broadcast of relevant news alerts and customised newsletters for the sub-brokers' clients.

Graphs for Sentiment Analysis

One of the most frequently asked questions is what analysts think of how people would behave in a certain scenario? In other words, people are interested in knowing what exerts think of a certain story and how people would react to the same. The company sees a great opportunity in publishing various sentiment graphs through its website and providing free widgets for its own publicity. The graphs could be specific to a specific company or for the overall index. It could be based on sentiments of experts or general public or a group of people coming from a specific region or with specific background, depending on what is supported by the TrendMiner technology.

Paid Seminars

HFPL's current objective is to make its clients and sub brokers aware of the objectives of the TrendMiner through free seminars and spread the word through mouth publicity. This is one of the reasons why the company had invited speakers from various exchanges at the investor awareness program. The long term goal of the company is to covert such free seminars into paid seminars where the seminars could be used for demonstration of the technology.

Collaboration with Stock Exchanges

Different exchanges offer different services. These include services such as providing company reports, company specific news and alerts, live stock market data feeds, delayed feeds, historical datasets etc. The company will, if possible, try to collaborate with relevant exchanges to sell the analysis in B2B market and generate income from the same.

Add-ons to Portfolio Management Services

The company itself trades in stock, commodity and forex markets. The analysis produced based on the TrendMiner technology would not only be helpful to the company to achieve better success in the markets but also be useful to manage portfolios of its clients who trust and rely on the expertise of analysts at the HFPL.

7.9 ONTOTEXT

For Ontotext TRENDMINER will bring the unique opportunity to collaborate with valuable partners such as two of the most prominent European language processing groups (DFKI and Sheffield) and Internet Memory – a leading European provider of web archiving platforms. In this way Ontotext will strengthen its position on the text analytics market with adoption of cutting edge technology and joint offerings.
In a market environment, where consumers value more and more control over the software solutions and vendor independence, selling product licenses becomes marginal and inefficient as a primary source of income. Increasingly our customers need software-as-a-service solutions and prefer to cut high initial costs for licenses and hardware. At the same time, they are ready to pay for tailored solutions and also usage fees according to quantitative measures such as amount of data processed, textual annotations, or queries.
In TRENDMINER, Ontotext develops a unique real-time, cloud-based text mining platform, which offers:

- Large scale real-time collection, aggregation, and storage of social media streams
- Real-time multilingual text processing on a cloud computing infrastructure
- Cross-lingual User Interfaces (UI) for browsing trends and sentiment in stream media

Such a platform will allow elasticity in the scale of our offerings and will become a pillar to our positioning on the international market for text analysis and semantic annotation solutions. It will also enable us to provide free and pay-per-use multilingual text mining services on the Web and thus establish Ontotext as a key European player in a market currently dominated by US-based OpenCalais and Extractiv services.

Bibliography

Maireder A, Ausserhofer J, Kittenberger A (2012): "Mapping the Austrian Political Twittersphere." In: *Proceedings of CeDem12 Conference for E-Democracy and Open Government*. 151–154. Danube University Krems

Maireder A (2011): *Links auf Twitter - Wie verweisen deutschsprachige Tweets auf Medieninhalte?* Vienna. http://www.univie.ac.at/publizistik/twitterstudie/ (last access 29.10.2012)

Grimes S, (2011): *Text/Content Analytics 2011: User Perspectives on Solutions and Providers.* Alta Plana

Fenn J, Raskino M, Gammage B (2009): *Hype Cycle for Enterprise Information Managemen*t. Gartner

Li C, Bernoff J (2011): *Groundswell, Expanded and Revised Edition: Winning in a World Transformed by Social Technologies.* Havard Business Press

Mislove A, Marcon M,Gummadi KP, Druschel P, Bhattacharjee B (2007): Measurement and Analysis of Online Social Networks. In: *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement.* 29-42, San Diego