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

D7.3 Application Final Results

Paul Ringler (SORA)

Abstract
FP7-ICT Strategic Targeted Research Project TrendMiner (No. 287863)
D7.3 Application Final Results (WP7)

In this deliverable we describe the final application developed for the WP7 Political Use Case in TrendMiner. This includes overviews of user requirements, development history, functions of the User Interface (UI) and an outline of future development plans.

Keyword list: Social Media, Natural Language Processing, Political Use Case, Prototype development

Nature: Prototype
Dissemination: PU
Contractual date of delivery: 31.10.2014
Actual date of delivery: 03.11.2014
Reviewed By: IPIPAN, DFKI
Web links: http://trendminer-project.eu/

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 Rvolution
F-93100 Montreuil, France
Contact person: France Lafarges
E-mail: contact@internetmemory.org

Eurokleis S.R.L.
Via Giorgio Baglivi, 3
Roma RM 0016, Italia
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
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
A-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

DAEDALUS - DATA, DECISIONS AND LANGUAGE, S. A.
C/ López de Hoyos 15, 3º, 28006 Madrid, Spain
Contact person: José Luis Martínez Fernández
Email: jmartinez@daedalus.es

Institute of Computer Science Polish Academy of Sciences
5 Jana Kazimierza Str.,Warsaw, Poland
Contact person: Maciej Ogrodniczuk
E-mail: Maciej.Ogrodniczuk@ipipan.waw.pl

Universidad Carlos III de Madrid
Av. Universidad, 30, 28911, Madrid, Spain
Contact person: Paloma Martínez Fernández
E-Mail: pmf@inf.uc3m.es

Research Institute for Linguistics of the Hungarian Academy of Sciences
Benczúr u. 33., H-1068 Budapest, Hungary
Contact person: Tamás Váradi
Email: varadi.tamas@nytud.mta.hu

Executive Summary

The project TrendMiner[1] aims at developing a social media monitoring platform for analysing user-authored content from social media, like Twitter and blogs, in order to provide better insights about emerging topics and trends and to improve the decision making process in different domains.

To this end a prototype social media monitoring platform in the domain of political analysis, political public relations and political journalism was developed. The functions of this prototype are intended to serve four groups of users: Politicians and party managers, journalists, PR experts and political researchers.

The prototype platform supports the following functions, which will be described in more detail in this deliverable:

- Real-time and historical analysis of streaming social media data
- Reliable identification of elements of political discourse
- Automated topic detection and summaries
- Quantitative content analysis
- Sentiment detection and analysis

In this deliverable we describe the final application developed for the Political Use Case in TrendMiner. This includes overviews of user requirements, development history, functions of the User Interface (UI) and an outline of future development plans.

[1] http://www.trendminer-project.eu/

Contents

1 Introduction

Politics is a volatile and complex business and in democratic modern societies that are steeped in different forms of media, political discourse happens in many different media and at a very fast pace.

The development of Social Media in recent years has also had a profound impact on politics. At the center of influential Social Media, Twitter stands out as attracting politically active and interested users, many of them politicians or professionals in the political sphere themselves (e.g. PR professionals, party officials, journalists).

It was therefore decided at an early junction in the project, to give Twitter a prioritized position as data source of for the TrendMiner prototype in the political use case, owing to the fact that the data produced by Twitter fulfils the criteria of being streaming data including also a dialogical component and which allows to gain statements about the users of this media, in addition of being a specific textual format that has seen much focus of research in the academic realms of both political sciences and computer sciences. Nevertheless, other data sources, such as blogs or news channels were always considered in addition to Twitter and provisions were made to include them at later stages of the development of both the TrendMiner platform and the political use case.

The development approach taken by the involved partners was goal oriented from the start: At every juncture of the project the focus was to set goals for the development of technologies for TrendMiner that would serve the purpose to build an open source infrastructure for analysis of political discourse to be adapted and developed onwards by others, empowering citizens, media and businesses in Europe.

First approximations of the results of the political use case of TrendMiner were already published as part of D7.1[2] and further developed to their present status quo during the course of the project. An internal discussion paper was circulated in 2012 and 2013 was a year of intensive work on the User Interface (UI) and functionalities.

Based on the research conducted by SORA, overall guidelines where formulated that guided the process of designing the final prototype for the political use case in WP7 of TrendMiner, but also informing work in the technical workpackages WP2, WP3, WP4 and in the integration workpacakge WP5.

Our main target groups consists of domain experts who are well positioned to supply the necessary context knowledge to be able to understand and interpret oblique references, innuendoes and to recognise most of the Topics, Persons and Organisations mentioned on Twitter, in Blogs or News Articles.

Our goal was to provide them with tools that will **assist them in their exploration and understanding of political discourse** by providing visualisations and forms of data aggregation to quickly gain an overview of what is happening before delving deeper into the data. Requirements and concepts for a UI for WP7, were realised and integrated as part of WP5. This work was also adapted for WP6.
Furthermore, work on analytical functions for WP7 to summarise content and detect trends over time was conducted throughout the project, as part of WP3 (clustering) and WP4 (summarisation).

[2] D7.1 Multilingual Public Spheres: Political Trends and Summaries, a non-public deliverable

Not only do our users posess the best possible knowledge about the political processes and actors that they are interested in. They also know what they are looking for, guided by the rythms of the daily news cycle and the occurence of political events such as election campaigns, parliamentary discussions or scandals. Even if users sometimes don´t know what they are interested in consciously, they will recognise it intuitively. We therefore want to **assist the user in identifying and interpreting relevant content** while leaving them to option to disregard content that they perceive as irrelevant if necessary. In WP7 a political ontology was conceptualised and developed, further work and integration was done under WP2 and WP4.

We aimed at structuring the resulting content in such a way, that our users are presented with **clear and stable categories, presented along a time axis and the possiblity to explore raw data, connections and sentiment**, to assist them in drawing their own conclusions. This was achieved as part of the visual analytics designed in WP7 and implemented under WP4, with the underlying political ontology as the structual "anchor" for analysis.

At the time of writing of this document, evaluation of the finalised prototype is underway as part of WP7 with results to be presented at the final review.

2 User Requirements

Social media are rapidly becoming an essential tool for those whose business is the analysis of politics and communication in the public sphere. At the same time they are drastically speeding up the news cycle, as platforms like Twitter emerge as an extremely fast medium of exchange for political discourse, breaking news and opinionated commentary of ongoing events.
On the level of users, three distinct groups of potential users of TrendMiner were identified with distinct sets of interests and requirements towards the TrendMiner User Interface (UI).

Professionals in politics and PR

This group was selected as main end-user group for the TrendMiner functionalities and technologies to be developed. Their needs and interests were therefore of central importance during development and testing phases.

Among them we find users such as politicians or party managers and PR professionals. This user group is highly active in their professional use of social media and highly aware of daily political discourse. Journalists should be considered to be part of this group as well, since they play an important role in politics by shaping public opinion as opinion leaders (although with lesser direct political influence). Here TrendMiner can serve as a quick and easy way of exploring trends and hot topics on Twitter or other social media, supporting this group’s professional needs.

Researchers

This is a group of users who need a powerful and reliable way of visualising streaming media, that also offers support for scientific research, i.e. providing means of making research reproducable, recordable and open to external review.

The general public

This heterogeneous group sums up interested citizens and political activists who want to monitor political discourse on a daily basis and who would approach TrendMiner from the same perspective as other, similar tools that are out there in the web. Their usage focuses on looking either at certain topics that they are interested in beforehand or explore trends and sentiment visualised by TrendMiner.

All groups overlap in some way, with some scientists simply browsing tweets on a certain topic for a few minutes or journalists and everyday internet users conducting deeper research into a story that has been unfolding on Twitter.

The target groups of TrendMiner all need to deal with social media data in some way: Politicians and party functionaries, journalists and PR-experts, as well as political analysts in the commercial and academic domain.

- All need a quick and easy way of exploring trends and hot topics on Twitter or other social media.
- Analysing and making sense of large quantities of textual data in a limited amount of time is a keystone in the work of all of these groups. This applies not only to textual data from social media, but also to other types of online texts, like online newspaper articles and blogs.
- Politicians, PR professionals and journalists need to be able to react quickly and flexibly to developing situations and the sooner a situation can be accurately assessed, the better their response can be.
- In the domain of academic and commercial political research, the analysis of historical data is somewhat more important and functions are required that support scientific research, making possible the replication of results, external review and data recording and sharing.

Based on our insights into modes of usage and user needs we formulated guidelines for the subsequent development of interface and analysis functions, which could be summed up as follwoing: "Don´t try to think for the user, rather support the user´s thinking and exploration process."

The software tool should therefore focus on what automated systems do best: Capturing content, ordering content, counting content and making content visible and accessible. One cornerstone is to **give the user the most relevant information at the right time**, taking advantage of the data processing infrastructure and data sources developed and processed as part of WP5 to set up a quick data delivery mechanism. This led to the development of nearly-realtime data analysis and visualisation mechanisms for Twitter and Blog data as part of the prototype for the political use case in WP7.

2.1 Preliminary Research

During year 1 (2012) and year 2 (2013) of the project, SORA conducted interviews with members of political parties and professionals from the public relation sector in Austria to flesh out the requirements of the protype in more detail. Below we present the key results from these interviews.

Twitter is an „Opinion cartell"

Based on their own observations, political professionals perceive Twitter as a place where public opinion is negotated and represents a new playing field for journalists and politicians. The number of individuals in these networks is usually limited. Personal acquaintance is the norm, rather than the exception, with interactions happening regularly at press conferences, interviews or short phone calls.

Due to this, one interviewee characterised Twitter as an "opinion cartell", where a select group meets online to discuss news items or voice their opinions on events and topics. Monitoring and participating in these discussions is viewed as highly important by both politicians, PR professionals and journalists. Politicians or their PR hope to be able to influence ongoing debates or raise new issues, while journalists view Twitter as a space where they can read and elicit public responses from political officials, thereby making it a fast and highly credible source of news.

Twitter impact on public opinion is generated by its publicity

The impact of Twitter on public opinion is generated through the powerful mix of two groups: Firstly, politicial decision makers who both use Twitter themselves and who become the focus of public discourse on Twitter. Secondly, opinion leaders such as journalists who use Twitter both to disseminate their views on political issues, politicians and their parties and to monitor what other opinion leaders say and consider important. The direct and easy accessibility of these public conversations, even by non-involved users, coupled with the possibility of getting immediate feedback to statements makes Twitter a big "global pub", where political discourse happens online, parallel to daily political work in parliaments, government ministries and other political players such as political parties, NGOs, interest groups, companies and citizen´s initiatives.

Twitter impact on public opinion is indirect

The impact of Twitter is considered by interviewees to be indirect, through agenda setting and the exchanges on viewpoints and issues that usually happen during the news cycle until the next edition of newspapers (both online and offline) is published. One interviewee described his impression of how journalists would essentially "post tomorrow´s headlines", giving political professionals a chance to react sooner and/or influence the content of the articles and newsitems in advance.

Twitter is extremely fast

The fast speed of information transfer that is common to most online applications is again enhanced on Twitter through the limit of 140 characters placed upon each Tweet. Thus, only the most important information is imparted, leading to a condensed and efficient transfer of information. Since Twitter is always online, always updated and always populated by users, this creates enormous pressure on political professionals, be they elected officials, party communications officers or their aides, towards being "online" constantly.

Twitter is highly dynamic, very attention driven

A side effect of the fast-paced information flow, coupled with the close acquaintanceship of most Twitter users in the political sphere is a tendency towards "hypes", where topics become more and more important, both due to being pushed and talked about by very influential Twitter users or by a great number of less influential individuals, creating a critical mass of Tweets.

Political actors anticipate and plan for increased use of social media

Social media is becoming an ever more important part of public discourse in general and political discourse in particular. Barack Obama´s successful election campaign in 2008 showed that Social Media can play an integral role in mobilising public support and voters.[3] The politicial professionals we talked to, agreed that Social Media would be playing ever more important roles in daily political discourse and election campaigns.

Professional Twitter use in daily communication work determined strongly by organisational structure & culture

The politicians and PR professionals we interviewed came from a variety of backgrounds, with differently sized and structured organisations. An important insight gleaned from the interviews was that the use of Twitter by professional communicators is always structured by the organisational structures and cultures that they are affiliated with. Smaller organisations or political parties tend to be more networked and egalitarian in their use, with many different actors working in loose coordination to introduce new topics or weigh in on ongoing conversations on Twitter.

Larger political parties tend more towards centralized policies, with briefings and guidelines on important topics being issued by dedicated officials.

Other political organisations who participate in political discourse on Twitter, i.e. social partners, ministries or companies weighing in on political discourse tend to reflect the level of centralisation and hierarchy inherent in their functioning. These are usually represented by a few individuals who reside at high or top levels of the hierarchy, e.g. ministers who use Twitter personally or their immediate aides such as press officers. This reflects the fact that the usual practices of coordination and approval of public statements further up the hierarchy are too slow and cumbersome in the face of the highly spontaneous and immediate nature of Twitter discourse. Extensive use of Twitter, beyond the simple reposting of messages an news-items therefore usually remains a prerogative of high-level leadership personnel who are empowered to participate in active exchange with other users.

[3] http://newsroom.cisco.com/feature/1006785/2012-The-Social-Media-Election- Retrieved on 02.10.2014
http://mprcenter.org/blog/2013/01/how-obama-won-the-social-media-battle-in-the-2012-presidential-campaign/ Retrieved on 02.10.2014

2.2 User needs

We inferred user needs both by asking direct questions on functions that a social media monitoring platform in the political domain should possess and from their description of their use of Twitter. The resulting points are:

(Very nearly) real-time data analysis and visualisation

Given the potential of journalists, politicians and other opinion leaders to react immediately to any relevant message spread on Twitter, timely information delivery is essential. Any useful tool therefore has to provide real-time updates and visualisations or a close approximation of this.

Linking of Topics and Users

Interviewees expressed a need to be able to engage with individual users who are involved in ongoing discussions about relevant topics. A function to connect topics and users is therefore necessary.

Information about the „news-cycle“,

In order to lighten the workload of constantly observing all ongoing discussions, a function is needed to assist in determining the ebb and flow of the news-cycle, to spot the right time to launch new topics, when existing discussion topics lose their attraction or when to be most vigilant in looking out for new topics popping up.

Aggregation and summarisation of information

In addition to supporting users by reducing the need for paying constant attention to Twitter, potential user express interest in functionalities that provide them with summaries and overviews of current and past topics without personally reading every recent and past Tweet.

Simple interfaces, the most important stuff needs to be easy to spot

Finally, politicians, PR professionals or journalists are in need of an interface that is easy to understand and doesn´t need to much prior skills or knowledge to learn its use.

2.3 Data sources

A fundamental part of TrendMiner´s goals was to develop an infrastructure for large-scale, cross-lingual Trend Mining and Summarisation of real-time data streams. A significan part of the project therefore revolved around identifying and accessing such data sources for the political use case.

For the political use case in WP7 Twitter was selected as the main data source and integrated in Year 2. Additionally Blogs and News Articles were selected as data sources for analysis and inclusion in the prototype. They share similarities insofar as they are textual media, which means they can be approached using the same method of analysis and visualisation. Their pub,ication speed and structure are quite different, however, so they are accessible as different sources of data in the prototype, so separate appraisal by the user is possible.

2.3.1 Twitter

Twitter was made the priority data source for integration, based on several reasons:

Relevance to political domain

As elucidated above, Twitter has rapidly become an important feature in the political domain, serving not only as a social network where opinion leaders exchange their views on a daily basis, but making these conversations public, thus making this influential kind of political discourse accessible to a wider audience than ever before.

In the case of WP7 this makes Twitter the prime source of up-to-date information and instrument of participation for the involved opinion leaders, for political analysts and for interested or politically active citizens. Recent events such as the "Aufschrei"-Debate that was started by several politically interested citizens and journalists[4] or the real-time visualisation of the Commission Candidates´ hearings in 2014[5] underscore this further.

Volume and velocity of data

Twitter is now considered to have over one 1 Billion users. While only 255 Million of them worldwide are active, this means that over 500 million Tweets are sent each day. By comparison, other sources of politically relevant information, such as online newspapers and blogs, let alone official sources such as ministries or parliaments offer much less daily volume and velocity. Based on statistics collected through the webcrawl created for WP7, of around 200 Sources covered, 50 % only add new content every two weeks, around 20 % add new content only every 20 weeks or so. This is contrasted with between 3000 and 5000 daily Tweets on average that are processed through the account set up for TrendMiner, which monitors the Austrian Twittersphere.

For WP7, the focus here was on creating a prototype to visualise political discourse (WP4 and WP5), capture it by engaging the entitiy detection capablities developed as part of WP 2 and analysing it using sentiment detection capabilities and advanced analytics developed in WP2, WP3 and WP4 respectively.

Multilinguality

Twitter is used by Users speaking 61 languages worldwide. The TrendMiner project now covers English, German, Italian, Polish, Spanish and Hindi and Bulgarian and all of these languages are also present on Twitter. For the purpose of WP7, the focus was placed on Austrian Tweets, i.e. Tweets in German, but using the multilingual resources developed as part of WP2, this coverage can be easily expanded to other languages.

Data structure

The structure of Twitter data is highly utilitarian and was designed from the ground up to facilitate further processing and analysis, using the *.json format to define clear sections of meta-data on the authors themselves, time and date of sending,

[4] https://twitter.com/Tugendfurie/status/294587569340022785 Retrieved on 02.10.2015
http://www.spiegel.de/panorama/gesellschaft/aufschrei-interview-zur-sexismus-debatte-auf-twitter-a-879729.htm Retrieved on 02.10.2015

[5] http://ephearings2014.eu/epdashboard Retrieved on 02.10.2015

geographical information, text body, conversations etc. Furthermore, the API provided by Twitter allows follow up analyses which may be prompted by social media monitoring tools such as the TrendMiner prototype.

Twitter Data Collection

For the purposes of building the prototype and developing the technological framework for TrendMiner, a specialised Twitter account was set up by SORA that monitors the Twitter users of the Austrian political Twittersphere. The composition of this account is based on academic research by Austrian media researchers (Maireder 2011, Maireder et al 2012), lists of opinion leaders made available to the public by the Austrian Press Agency[6] and constant human curation by SORA analysts to ensure that the users tracked are indeed highly engaged in commenting on political discourse, by screening profiles and timelines. The end result tracks the 2000 most relevant Twitter users, who are highly interconnected: Collectively these users have published over 8,7 Million Tweets over the lifetimes of their accounts. On average, each user on this list has 1391 followers, with 50 % having more than 449 followers. The average number of users they themselves follow is 526, with 50 % following at least 353 accounts[7].

Further analytics of the SORA account[8] show that this selection of accounts, while not very large, is of very high quality. Most accounts are well established (over 70 % are older than 4 years), and highly active, with 63 % of all accounts having published at least one Tweet in the last Week. 54 % of these accounts have written more than 1.000 Tweets in their lifeime, and only 1 % highly active accounts with more than 50.000 Tweets. Furthermore the proportion of accounts that issue mainly tweets and retweets countaining URLs is low: Of the accounts that are tapped for data by TrendMiner, 28 % have a proportion of more than 50 % of Tweets containing URLS and only 9 % have more than 50 % Retweets. This means that the user base covered here is actually commenting on events, and not simply tweeting news items.

Table 1. Account ages of users in SORA Account

Account ages	
Age	**% of Accounts**
1 - 12 months	1%
1 - 2 years	8%
2 - 3 years	21%
4 years +	70%

Table 2. Recent activity level of users in SORA Account

Recent activity level	
Time since last Tweet	**% of Accounts**
1 day ago	39%
1 week ago	24%
1 month ago	13%
1 - 6 months ago	12%
> 6 months ago	12%

[6] http://twitterlist.ots.at/

[7] Based on own calculations

[8] Source: Followerwonk.com

Table 3. Total no. of Tweets of users in SORA Account

Total activity	
Total no. of Tweets	**% of Accounts**
0-1k	46%
1k-5k	41%
5k-50k	12%
>50k	1%

Table 4. % Tweets containing URLs per Users in SORA Account

% of Tweets containing URLs per User	
% of Tweets per User	**% of Accounts**
0 - 25 %	50%
26-50 %	23%
51-75 %	11%
76-100 %	17%

Table 5. % of Retweets containing URLs per User

% of Retweets containing URLs per User	
% of Tweets in SORA-Timeline	**% of Accounts**
0 - 25 %	68%
26-50 %	23%
51-75 %	7%
76-100 %	2%

In Year 3 (2014) of the project, further extension of the prototype language capabilities was achieved for the political use case through the inclusion to Polish and Hungarian as part of WP9 and WP10. Deliverable D10.1 describes the new components implemented by the new partners.

2.3.2 Blogs & News Articles

Blogs and news articles represent another facet of online political discourse. For Austria alone there we estimate there to be at least 300 Blogs dedicated to discussing political topics, in addition to around a dozen dedicated news portals. While this represents a large reservoir of data sources, each one is an individual source which needs to be vetted and which requires automated data collection to be adapted separately to allow proper analysis in conjunction with Twitter data.

Blogs and News data documents share a similar structure insofar as each published document consists of 3 discrete parts: title, article body and comments. Each discrete part may contain one or more references to political entities and politically relevant topics, which may also overlap. Title and text body usually refer to a limited number of political entities and topics, which all share clear relationships that are established by the author. The comment section is more varied, with commenters usually coming from very different backgrounds and different motivations for commenting. Comments may or may not be curated by a moderator and therefore contain varied percentages or spam. They differ however in the way the data is embedded in the HTML-structure of the Web, leading to "pollution" of processed textual data by other textual parts of the website in which title, text body and comments are embedded in.

As part of the work for TrendMiner, the partner IMR has created a crawl of Austrian Blogs and News portals, based on a list of 91 Hyperlinks created by SORA. It consists of hyperlinks to different Newspapers and Websites that carry political news and commentary. These Hyperlinks were been annotated with Country names, Language, Domain and scope. An additional list of the 65 most important Blogs was also researched and created in for this purpose in 2014 and used for IMR to refine their search.

The goal was to provide a clean feed of data containing descrete data on document identifiers, authors, time and date of publishing as well as article title, text body and comments data.

2.3.3 Additional data sources collected

In the course of the project, a variety of additional data sources were screened and collected as part of WP7, an ongoing process throughout the project. These data were used at many junctures during the project, providing insights and gold standard data which formed the basis for development of prototype functionalities and analytical functions. As part of the extension of the project consortium in the last year, additional data sources have been collected for the new languages, Polish and Hungarian. This is described in the deliverables D5.5, D9.1 and D10.1

Polling data

An extensive collection of polling data from Austria and the UK (see also see D3.1.2) was compiled, partly during 2012 and 2013, but also during 2014.

2010:

- 179 UK-wide polls between January and May, leading up the the UK general elections on May 6th 2010, collected by SORA.
- 12 Vienna-wide polls between August and October, leading up to the city council elections on Oct. 10th 2010, collected by SORA.

2012-2014:

- 84 Austria-wide polls between September 2012 and November 2013 until the general elections on Sept. 29th 2013 an somewhat beyond, collected by SORA.
- 16 Austria-wide polls between November 2013 and May 2014, leading up to the European Elections on May 25th 2014, collected by SORA.

News Articles

A collection of corpora containing political news articles was compiled during 2012 and 2013:

- EU News Summary Text Corpus: This corpus consists of the press summaries of EU news made by the think tank Open Europe and covers EU related topics between 2006 and 2012[9]. Compiled by SORA.

- Living Knowledge Data Corpus: Sentiment-loaded political texts collected during the LivingKnowledge[10] project. Compiled by SORA.

[9] http://www.openeurope.org.uk/Page/PressSummary/en/LIVE

- Reuters News Corpus: Large, manually categorized corpus of newstexts, available from Reuters[11]. Resarched by SORA.

2.3.4 Annotations

Part of WP7 was dedicated to the creation of gold standard data. Several such corpora were created by SORA between 2011 and 2013:

- Accounts Identifiers for Austrian Twitter Users, based on publicly available lists and further refined.

- 40 tweets, extracted from a corpus of tweets about the elections in Vienna in October 2010, collected via the Twitter Gardenhose. This corpus was used to test the LODIE tool of the partner USFD (WP2) and develop the annotation scheme, particularly informing the development of the political ontology (WP2).

- 200 Tweets were extracted from the SORA account, covering the period from Dec 1st 2012 until Jan 16th 2013, intended to cover a short, but intensive discussion was held among Twitter users on topics like democracy, censorship, culture, sexism, racism, surrounding a scandal in Vienna. It provided insights into sentiment and opinions on Twitter and informed further conceptual work.

- Gold standard Annotations for Summarisation for WP4. This involved the creation of tweet filters using a tag cloud, and then the annotation of up to 8 tweets relevant to those filters as interesting or not interesting. This task was carried out in real time on contemporary data, and the results were used to compare many tweet summarisation algorithms automatically using ROUGE and MAP (See deliverable 4.3.1). In total 60 tweet sets were created with gold standard annotations.

- Further, more general Annotations for USFD for Persons, Locations, Organisations, Products (1200 Tweets) on Crowdflower

- Following the county elections in Lower Austria and Carinthia in March 2013, a corpus of 2640 related Tweets from Feb 17th until Mar 1st was extracted from the SORA account. 379 Tweets were annotated for sentiment towards political parties, leading candidates and TV reports, all relating to the Lower Austrian elections. In addition some Tweets related also to the Carinthian elections were annotated in the same fashion.

New and updated annotations in 2014 include:

- Manual filter creation: Historical data was revisited and new collections of keywords produced for the purpose of supporting manual evaluation without

[10] LivingKnowledge was funded under the 7th Framework Programme (Project No. 231126), http://livingknowledge.europarchive.org/

[11] http://trec.nist.gov/data/reuters/reuters.html

reusing filters from the gold standard. Another 17 tweet sets were created in this way, this time without any existing relevance annotations.

- Manual evaluation: The created 17 filters were used to generate sets of candidate tweets, and these were summarised automatically using the best performing system and several baselines. SORA annotators then considered each summary and rated them on the basis of utility, redundancy, usefulness and subjective preference. These scores were used to more rigorously compare 5 approaches to summarising tweets.
- 15 sets of gold standard summarisations, containing 50 Tweets each were created in September 2014, in order to corroborate the results from manual evaluation, which differed with those of automated evaluation. The gold standard annotation task was carried out once again, with a selection of 15 tweet sets from the original run. These new annotations were compared with the originals to see if the information needs of SORA had changed in the intervening time. Only weak agreement was found between contemporary SORA and SORA half way through the project. More research would be needed to fully understand this variation.

- 400 clusters based on the EU News Summaries Corpus, produced as part of WP4, were annotated for topical, spatial and temporal coherence in April 2014.

- 400 clusters, based on tweets collected by USFD from June 2012 to June 2013 from 37 different cities in Austria and Germany as part of WP4 were annotated for topical, spatial and temporal coherence in March 2014.

- Gold standard data was collected through the SORA Twitter account, drawing on the "Twitterbarometer" experiment, where hashtags of political parties were used to denote positive and negative sentiment.

- A list of important Austrian political figures created in 2013 and updated throughout 2014: Austrian Candidates for the General Elections 2013, Austrian Candidates for the Elections to the European Parliament in 2014, Members of the Austrian National Council, leading officials of major Austrian political parties, members of the Austrian federal government, leading officials of Austrian interest organisations (Worker´s Chamber, Chamber of Commerce, Chamber of Industry, Workers´ Unions). Also expanded and adapted by DFKI, IPIPAN and RILMTA.

2.4 An Ontology of the Political Domain

One fundamental task of the TrendMiner prototype developed during this project was the structured presentation of the content of political discourse on Twitter.

2.4.1 Capturing politically relevant content

This requires a framework for identifying political discourse, i.e. an ontology of the political domain under consideration in Austria. A political ontology organises political content into classes and relations. One of the most important requirements

for a tool covering real-time political discourse is that it has to provide a concise summary and overview of current and past political discourse. One important way is to provide a systematic overview of relevant content, structured so that the user can quickly navigate his or her way through the mass of constantly streaming data .

The main challenge in all of this is, that while certain elements of political discourse stay relatively constant over time (e.g. names of political parties and other organisations, names of political figures, names of political processes, names of political regions etc.), others are in a state of constant flux, especially topics of political discourse. Here both signifier (i.e. names) and the signified (i.e. what those names refer to) may change over night, as events (political or otherwise) happen and are discussed over social and other media.

The goal here was therefore to develop an ontology that would cover as many stable elements of discourse as possible, while excluding unstable elements. The main challenge of analysis then becomes the task of automatically identifying unstable elements (without integrating them into the ontology) and visualing their relationship with the stable elements.

Entities or topics, even words, can be political by themselves, such as “Law” or “Tax” or “Parliament. Political events such as elections are by definition political and include a host of related entities, topics and terms that are politically relevant.

Other terms or entities are made “political” by their context, by being mentioned as part of a with political discourse, e.g. being the focus or subject of a political process. “Coffee” can be tweeted about when you meet a friend, but also as apolitically relevant subject when tweeted about it in relation to Trade policy or discussions about globalisation. “Traffic” can be a nuisance that is tweeted about when commuting to work, or a politically relevant thing, once someone tweets that “they should introduce a law that forbids cyclists to use the roads because they are a nuisance. In an episode covered in an annotated corpus of Tweets in Trendminer, a rock band that is not recognisable as a political entity becomes one when it is the focus of a discussion about censorship, cultural politics and sexism.

A tweet or other document could thus be identified as politically relevant by checking whether an object is referred to in a political context, whether there is co-occurrence between this object and something contained in our political ontology. As mentioned above, this may also help to to identify relevant tweets in cases where ambiguous objects occur together, their co-occurrence making them clearly political or allowing a correct differentiation between named entities such as politicians, who may have similar names and be only distinguishable by the political context they belong to.

Research conducted before and during the initial phase of the project revealed only rudimentary resources available to reach these goals, especially when it comes to reaching the first 2 requirements. Therefore a political ontology was developed as part of WP2 and WP7.

In 2014 this Ontology was populated based on research by SORA and adopted and further expanded under WP2. It has has been made ready for integration into the prototype for WP7 and informs all features related to the extraction and visualisation of political content, i.e. entity search, visualisation, co-location detection and facetted browsing. The ontology has been updated by elements from the Polish and the

Hungarian political realities, and the updated ontology will be made available in December 2014 at www.dfki.de/lt/onto.

2.5 Sentiment detection

2.5.1 Political Sentiment: Emotions vs. Opinions

Politically Relevant Emotions

In TrendMiner we differentiate between the concepts of emotions and opinions when dealing with political sentiment.

In recent years, disciplines like Linguistics, Political Science and (Neuro-)Psychology have drawn their attention to emotions as crucial factor for understanding political processes and outcomes (Hofinger & Manz-Christ 2011). In classical voter studies (Lazarsfeld et al 1944, Campell et al 1960), emotional attachments explained why voters usually were not inclined to change their preferences. Choice, especially in the context of changing your preference, was seen rational, the outcome of a cognitive process (Downs 1957, Fiorina 1978), as some noted, however, only on a partially informed base (Popkin 1991).

Today we know that choices and human decisions always rely on an interplay of cognitive and emotional instances of our minds (Damásio & Damásio 1996), and that emotions are in the very core of our political behaviour. Our (emotional) brain is "built for Politics" (Schreiber & Fonzo 2012). Emotion is therefore seen to an important role in politics, because they act as basic driving forces behind many political phenomena and trends.

What are the emotions that influence voters the most? According to Brader (2008), the most effective emotions in political campaigns are enthusiasm, anger and fear. This has been experimentally proven for political advertising. which is only one form of political communication. Especially in case of TV commercials, there a variety of dramaturgical tricks can elicit emotions (music, background noises, imagery, the tone of the speaking voices).

There are scientific efforts to measure emotions from text in general and Twitter in particular. These efforts rely on identifying emotions as expressed in tweets through human annotators. The defintions of emotions used and the scope of emotions researched are diverse. To the best of our knowledge many (Kirk et al 2012, Valitutti & Stapparava 2004, Strapparava & Mihalcea 2008, Mohammad 2012, Inkpen et al 2009, Bhowmick et al 2010, Alm et al (2005) are based on Ekman's (1972) basic set of six emotions anger, disgust, fear, happiness, sadness, and surprise or derivations thereof, as it provides a semantically distinct set of emotions to identify an is furthermore applied in popular lexical resources such as SentiWordNet (Strapparava & Valitutti 2004). Others use simplified definitions (e.g. Chmiel et al 2010).

While the theory underlying the significance of emotions in politics is sound and widely recognised, emotions still remain purely internal states. Their expression is hard to measure exactly by human readers or through automated methods, using mostly textual media like Twitter, news articles or blogs.

Political Opinions

Political opinions have a direct equivalent in human speech. Unlike emotions, which are much more diffuse, opinions can be identified through syntactic reasoning and attributed to opinion holder and objects of opininion. They can be voiced implicitly and explicity and are a typical feature of political discussions in any medium, including text-only media such as Twitter, News articles or blogs. Opinions are usually defined as having positive, neutral and negative polarity in text and identified through using libraries of terms that receive a polarity value. These techniques are well established (Bethard et al 2004, 2006, Kim et al 2006,) and have received constant attention in academic NLP research (O'Keefe 2013, Bing 2012, Wiegand & Klakow 2011, Bo & Lee 2008, Maynard et al 2012, Maynard & Funk 2012).

As part of WP7 a schema was defined which is taken to fit the political use case best:

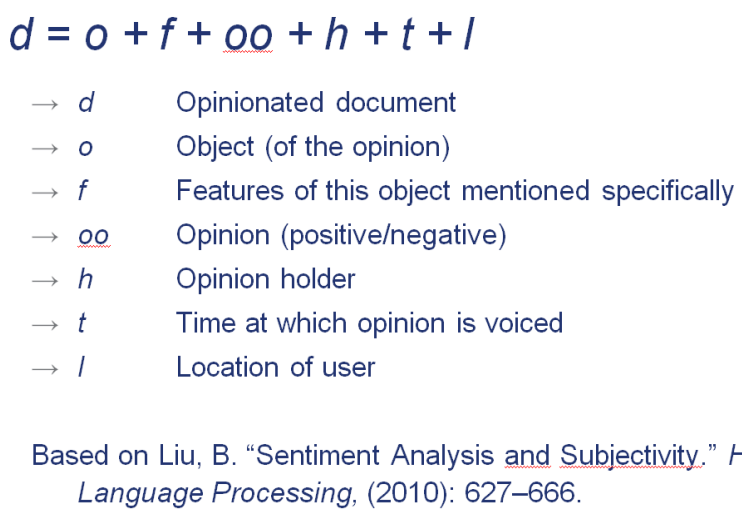


Figure 1. Political opinion analysis scheme

Opinionated documents should be processed to yield information about the object of the opinion, together with any features of that object that were mentioned. The opinion itself is taken to be either of positive or negative valence with relation to the object of the opinion or one of its features. In addition to a time stamp, the holder of the opinion (i.e. the author of the document), as well as the relevant location should be processed and visualised. This formula, while simple, has the advantage that it can cover both Twitter data and Blogs and New Articles. It can also cover other data types that may be included in the future. All the features named above are included in the opinion module of the TrendMiner ontologies, and correspondingly annotated tweets or other textual sources are populating the ontology.

In the case of Trendminer, the basic positive-neutral-negative polarity scheme is suitable for application to political figures, such as Politicians, Parties and Political Organisations. Another relevant instance is political deliberation, where discourse may revolve around a certain political measure and whether it should be carried out or not.

Facet-based opinions are a particular feature of analysis which can be applied to political discourse, mainly to political actors. As part of the political ontology, these can be defined to posess attributes such as "conduct", "success" or "competence". Providing real-time information about these attributes for relevant players on the

political field would add an important layer of information which helps the users to understand political discourse on Twitter, Blogs and News Articles.

As is the case of the political ontology, research during 2014 uncovered a lack of domain specific resources which can be easily matched to the political ontology developed until now. The first version of the political ontology is available at www.dfki.de/lt/onto.

A particular challenge in detecting political opinions lies in the fact that in political discourse statements are always in some ways connected to political values, or more generally, ideological stances. Depending on the ideological stances of both the opinion holder and the object of the opinion, polarity of sentiment may differ wildly. By the standard of a person with a liberal democratic stance as is prevalent in Western Europe the word "authoritarian" is considered to have very negative implications while a supporter of an authoritarian regime might not see this as the case. Ideology may also influence individual stances towards issues.
There is current research on how to predict the political stance of Twitter users and other speakers (Himelboim et al 2013, Demartini 2011).

2.6 Analytical Functions

Part of the goals of TrendMiner is the development of new methods of detecting topics and trends, particularly by bringing statistical and symbolic analysis methods together. For WP7 several such methods were developed, based on the requirements analysis of the users.

2.6.1 Basic Statistical Indicators and Visualisations

One goal of TrendMiner is to assist the user in generating knowledge about trends, events and topics that occur in social media. The traditional way of assisting analysts in making sense of large datasets is to break them down into numerical statistics, using quantitative content analysis methods and corresponding visualisations.

Numbers of mentions

For the TrendMiner prototype "Numbers of Mentions" was chosen as the main numerical indicator, because it offers the best insight into the appearance of a topic on the "radar screen" of the public, so to speak. It is also applicable to any kind of document that TrendMiner may process, be they Tweets, Blogs or News items and can be combined with a number of other statistics to form new indicators. Numbers of mentions can be comined with the authors of Tweets to form. Sentiment values are associated to the mentions on the base of key words detected in their surroundings.

Possible additional basic statistical indicators:

The software framework of TrendMiner could be expanded in the future to include additional indicators,

Indicators may give information on how frequently a term (e.g. a person or a hashtag) is mentioned, which other terms it co-occurs with, the numbers of “likes”, “retweets/shares” or “replies” its mention is associated with, and so on. Further possible and useful indicators refer to the exposure that Tweet or any other document has received.

User-Metadata can in general play an important role and be combined to the statistical analysis of the text. Example of such metadata are:

- Type (Categorical, e.g.: Austrian Media, Austrian NGOs, Austrian Bloggers and other Influencers, Political parties and other political organisation, Journalists, Politicians)
- Location
- No. of Tweets (numerical)
- No. of Followers (numerical)
- No. of Following (numerical)

Maybe other, automatically generated data

- Activities like No. of tweets/replies/Retweets per hour (numerical)

2.6.2 Advanced textual analytical functions

Basic quantitative indicators are indeed useful on a very abstract level of analysis, e.g. for determining especially popular topics among a certain target group, or monitoring events live.

However, they remove the data from its original, textual state, thus making an understanding of the more intuitive or "soft" aspects in the data impossible. This is even more important considering the difficulty in attaining valid results from current methods for automated sentiment analysis.

Also, automated statistical analysis methods predetermine analysis outcomes in some ways, because they are limited to what the algorithms and databases where built to detect in the first place. This limits the possibility for the discovery of new, unexpected things.

Therefore it was decided to consider additional forms of advanced textual analytics which would provide other views on the content. Summarization and Clustering are two methods we have been investigating for presenting the data to the user.

2.6.3 Summarisation

In the case of streaming social media data, which is constantly changing, with new material being generated at every second, every tool that aids analysis, while preserving the textual nature of the underlying data, can be a significant bonus to users of social media monitoring applications who are managing communications and public relations. These jobs can require quick and flexible reactions to developing situations. The sooner a situation can be accurately assessed, the better the response can be. Traditional numerical indicators can give accurate and valid information about the public attention directed towards certain topics, but in combination with automated summaries, these numbers gain "life" and allow pr professionals to engage their intuitive faculties.

In the case of more scientific applications, automated summaries are step towards speeding up a time-consuming, yet necessary task: The analysis of textual data is a keystone in many projects in the domains of social science and political analysis. This applies not only to textual data from social media, but also to texts generated by other means in the course of research, like interview transcripts or other documents that may constitute research data. Particularly commercial research firms, such as SORA, and their clients, stand to profit from a tool that saves on time and thus, costs.

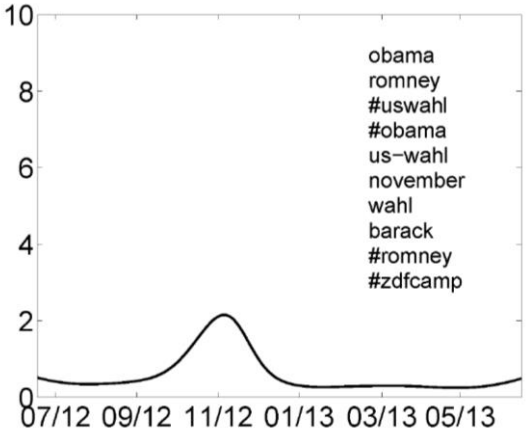
By displaying 3 to 5 real, existing documents or other texts that most closely represent the other thousands of documents that are being generated simultaneously, analysts receive another tool to quickly get a “feel” for developing situations in the case of real-time data, or for historical data, without having to reconstruct its contents from abstract numerical statistics or reading through an incomplete and arbitrarily chosen subset of documents which may or may not be a representative sample. In combination with tradition quantitative content analysis methods this is a useful addition to the analyst´s toolbox.

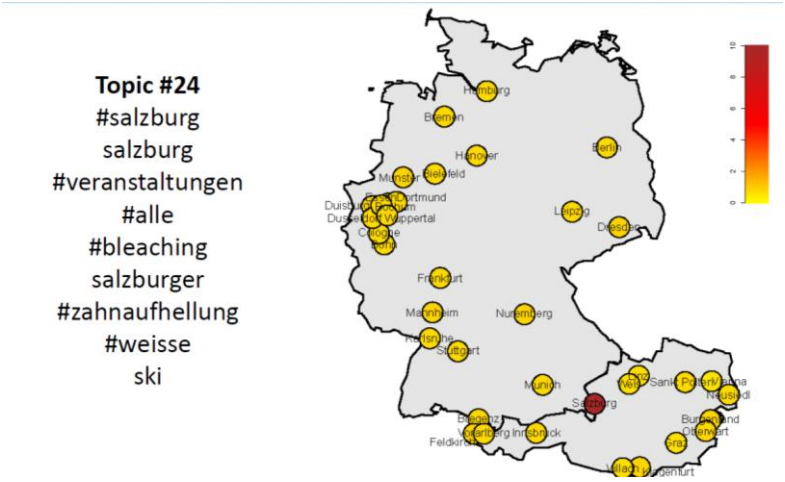
The summarisation tool and its related visualisation developed therfore withn WP4 is described in deliverable D4.2.1.

2.6.4 Clustering

WP7 also integrated the spectral cluster algorithm developed in WP3 (see deliverable 3.2.1 for details). Results of the clustering algorithm applied to tweets dealing with political themes are now presented to the user in the form of word clouds. Current work is dedicated in combining the terms included in the word cloud in more sensitive relations.
This work corresponds to an attempt at making trends visible in retrospective. Future applications of this may involve using these to predict future clusters on topics based on the information on similar "historical" clusters. This requires more research into the behavioral patterns of attention and news cycles and the interactions between social and traditional media such as News papers, Television, etc. The process is described just below, and we show also some current visualisation possibilities

- find `topics’ in a collection of documents (tweets)
- each document is assigned to a few `topics’
- `topic’ = a set of semantically coherent words
- can include temporal dependencies and evolution





2.6.5 Political Forecasting

As part of WP7 political opinion polls were collected and used to assess the possibility of using such data to forecast of election results. This technique has already been applied in the past, e.g. by Tumasjan (2010) or Metaxas (2011). Gayo-Avello (2012) later performed a meta-analysis of past studies showing that prediction power was lacking. As part of the TrendMiner project, similar approaches have been tried and lead, contrary to the conclusions by Gayo-Avello, to promising results (Lampos 2012, Lampos et al 2013).

2.7 User Interface design and development

The purpose of the TrendMiner UI is to allow users to explore the twitterverse or other social media in new ways, to "see" or "feel" their way around the semantic world that is the web 2.0. Development aimed at reaching the following standards:

- Visually attractive
- Self-explanatory
- Informative
- Dynamic & real-time
- Basic Explorative functions

The TrendMiner prototype needs to be an easily accessible tool that requires very little from users in terms of hard- or software and can be widely publicised with minimal effort. It is directed at specifically at professionals in the political realm, but should also be accessible for users from the general public.

TrendMiner should also support empirical research in the social and politicial sciences. Additionally, the interface should support the following fundamental requirements of scientific practice:

- Replication
- External review
- Data recording and sharing

User Interface design and development proceeded in several iterative stages and was also tied to the development and implementation of analytical and visual functions. Each stage was followed by an examination of results and re-evaluation of requirements.

Stage 1: Requirements definition & rough mockups

Stage 1 of development began once research into User requirement was complete. It consisted of workshops with Partners SORA, ONTO and EK held in February and July 2013. Starting from early sketches (e.g. Figure 2. Early UI sketch), the interface was developed further, always taking into account the additional analytical function which would be integrated at later stages.

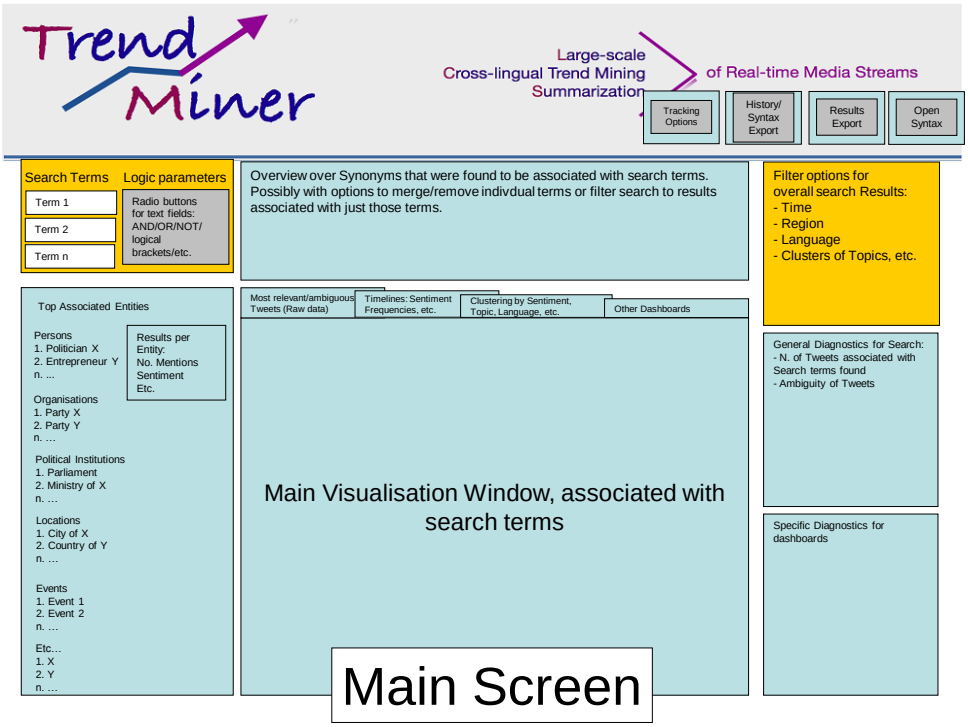


Figure 2. Early UI sketch

Stage 2: Static data prototype

A static data prototype was produced in October 2013. This prototype already fulfilled many functions of the envisioned goal and was presented at the 2nd review meeting.

Stage 3: Real-time data prototype

Real-time data analysis and analytical functionalities where added during 2014. This phase was characterised by intensive internal tests of the interface, as many issues with functionalities, entity definition, sentiment analysis and data sources had to be resolved in other workpackages. following comprehensive initial feedback on bugs and functionalities of the static interface.

3 The final prototype version of TrendMiner

3.1 The "Tracks" page - Initial, persistent filter settings

The finalised structure of the UI is based on a structure of "Tracks" (Figure 3. Track configuration) that allow the user to restrict the vast amount of documents flowing into the system by defining a starting point of interest. The "Track" should be viewed as the overall area of interest an is best defined with a specific, but broad area of political discourse in mind. This could e.g. be a generalised track that monitors all tweets, news or blog entries associated with political parties in a country posted in the last hour, or that allows for the tracking of a specific political event, tied to specific political entities, such as persons or institutions, e.g. by specifying incumbent candidates in an election campaign.

Several search dimensions are provided for this task:

- Topics and mentions
- Region of origin
- Data Sources
- Language of the resources
- Time period

This provides users with a flexible starting point from which to explore and analyse politicial discourse as it happens.

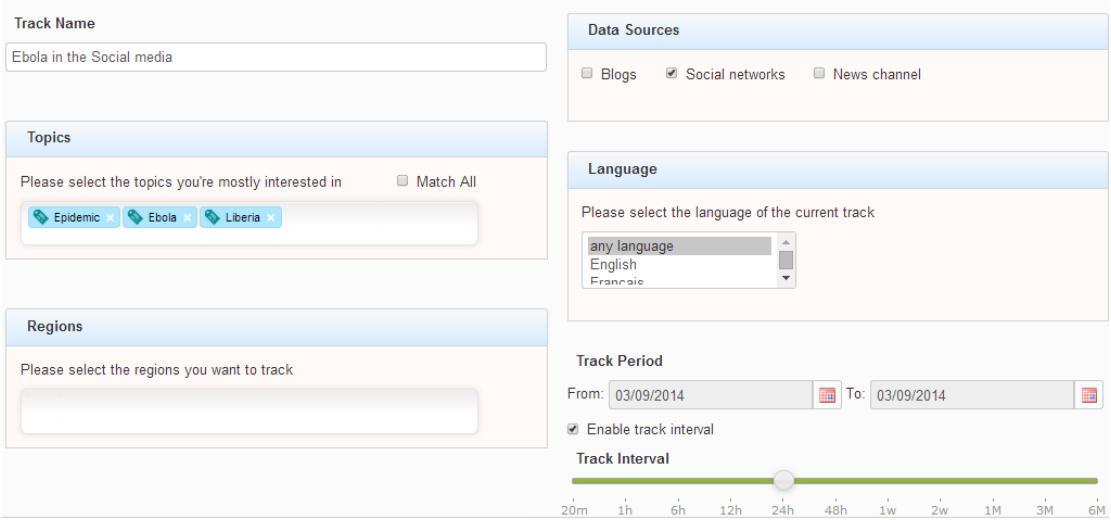


Figure 3. Track configuration

Topics

Here the user enters the names of the actors and other entities of political discourse. These can be persons such as politicians or other public figures, they can be political organisations, such as political parties, ministries, NGOs and other organisations that relevant in politicis such as, such as media or business companies.

The basic function engaged here is the tracking of mentions of these actors and objects as well as those associated with them.

Region of origin
TrendMiner is a multilingual tool and as such can track public discourse across borders. This function is designed to help the user limit search to political discourse that is related to certain geographical regions, using the geotagging function of Twitter. It is possible to define either geographical or political regions.

Data Sources
In its fully finished stage of develpment, the TrendMiner tool is capable of accessing and processing not only Twitter as a social media source, but also Blogs and News channels. These Data sources are displayed and analysed separately from Twitter data, as they follow a different, slower news cycle, and as such, numbers of mentions of objects have a different kind of significance here than on Twitter.

Language of the resources
Another feature to help the user limit their field of view on the data is the language selection. Politics is a social process is always connected to national context and therefore language is an important means of narrowing searches so that results make sense.

Time period
In a fast paced social area, such as politics, the timeframe of observation is a crucial variable. This dimension allows users to select between long-term historical analysis stretching over several months and short term analysis that offers nearly real-time updates, refreshing data every 20 minutes.

3.2 The "Charts" page - Displaying search data

Following the definition of a track, the user is presented with the "Charts" interface. This interface presents the user with the results of their track in two different ways, always taking into account further restrictions put into place by the user, such as the timeframe, geographical location and language.

The central result are always the policital entities the user is interested in. Information is structured such that the user can understand how often the entities they are interested in where mentioned over specific timespan (Figure 4. Tracked Entities along timeframe).

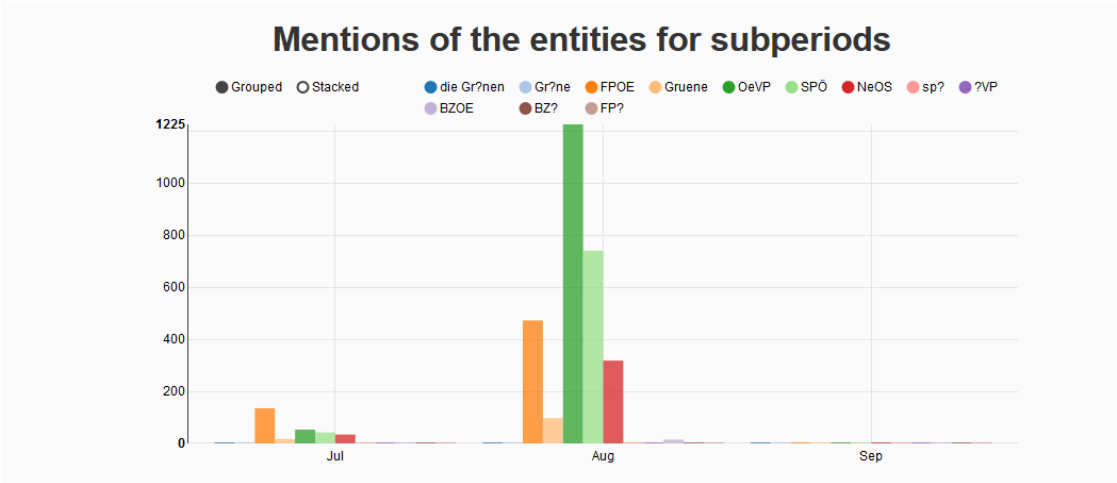


Figure 4. Tracked Entities along timeframe

The corresponding levels of sentiment that were noted in the data for the same time period are located directly below (Figure 5. Average sentiment along timeframe). This helps to gauge the overall direction of Sentiment of the debates associated with the political entities mentioned.

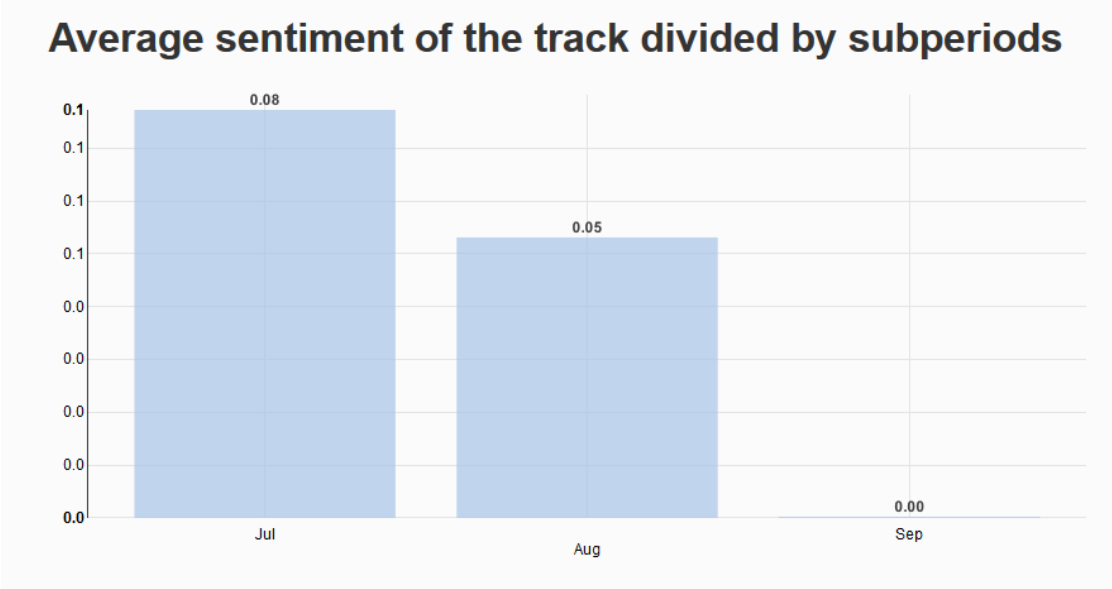


Figure 5. Average sentiment along timeframe

Further down, the geo-location tags contained in the data are visualised on a map (Figure 6. Geolocation tags displayed on map), helping the user to localise the tweets in the data, therefore lending a spatial context to the discourse visualised.

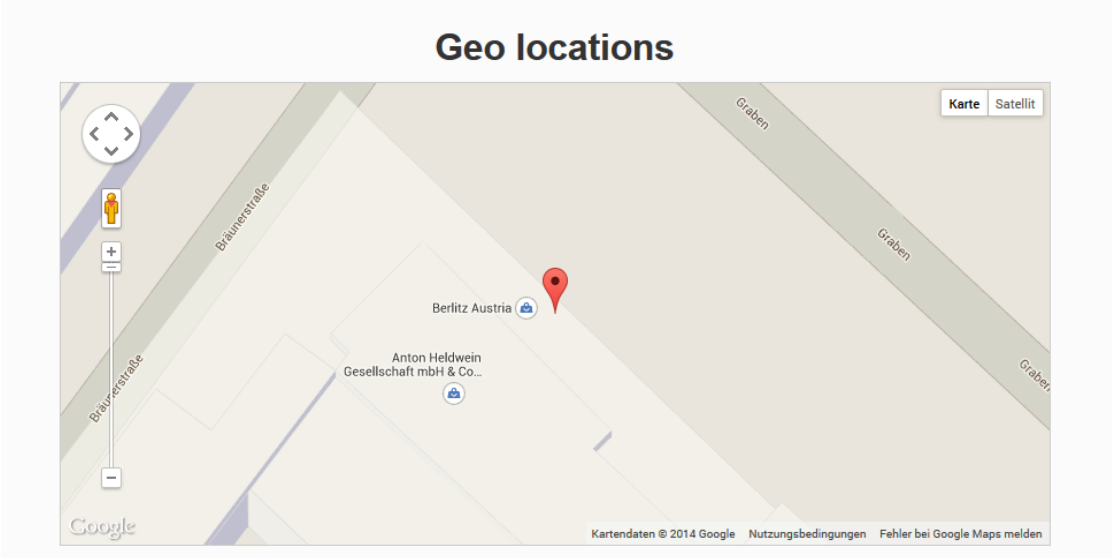


Figure 6. Geolocation tags displayed on map

Finally, the co-occurence of entities (Figure 7. Co-occurence of entities) is visualised on a matrix at the foot of the screen. This provides the user with an overview of the connections between the entities they are interested in and other entities that they are mentioned together, immedeately establishing the context in which they should be viewed, giving valuable information about political discourses as they happen.

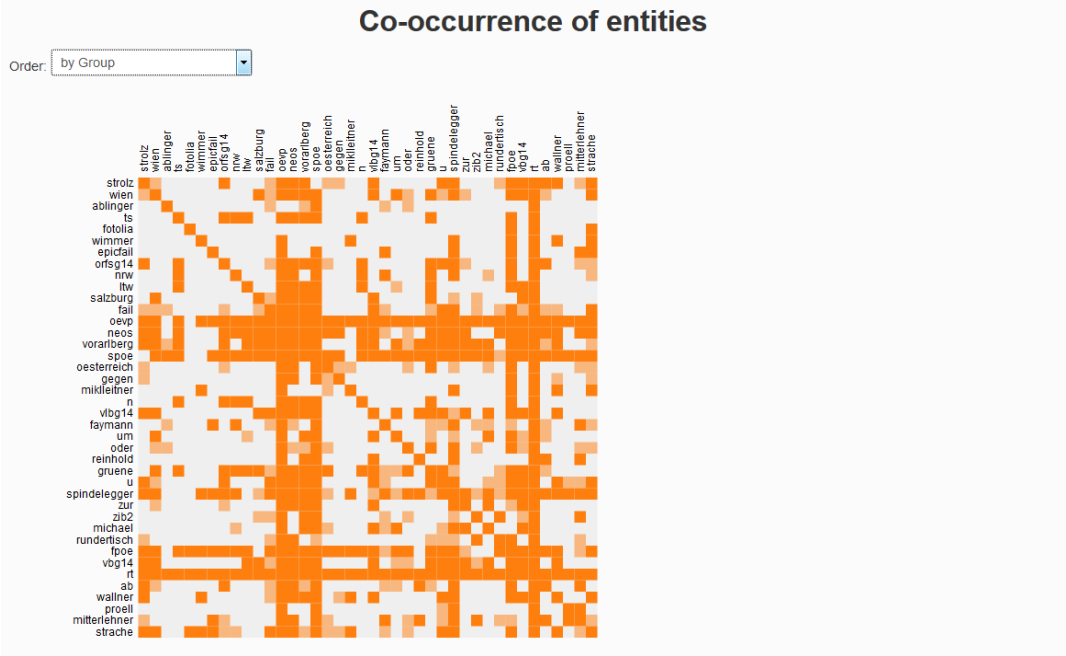


Figure 7. Co-occurence of entities

3.3 The "List" interface - Raw data

The "List" interface displays the raw data that makes up the charts page. All documents, be they Tweets, Blog entries or News items can be viewed here (Figure 8. Raw data browser).

Figure 8. Raw data browser

3.4 The "Filter" page - Further refinement of searches

To facilitate data exploration and interpretation, the user is given the option of refining the results of the track they specified. The "Filter"-page has a two-fold function: On one hand it acts as a filter for the initial track results, allowing the user further narrow down the displayed data, to display only the most relevant sections. On the other hand it acts as a textual content display, offering overviews and summaries of the political discourse.

The first item on this screen is the time-axis filter (Figure 9. Further filtering along the time-axis). It consists of an overall view of the numbers of entities defined in the Track settings, analogous to the "Charts" page, albeit with a more finely spaced time-scale. This is supplemented with a slider-function to narrow down the timeframe. The higher granularity of this scale means that it functions both as a detailed display of data and a filter. It allows the user to quickly spot interesting stretches of time, eg. rising trends, spikes and lulls. They can then narrow search results down to the most relevant stretches of time, e.g. excluding periods with few or now tweets at all (especially important for analyses that cover only 24h or less, where night-time lows in activity can distort the overall result), or highlighting stretches with relevant patterns.

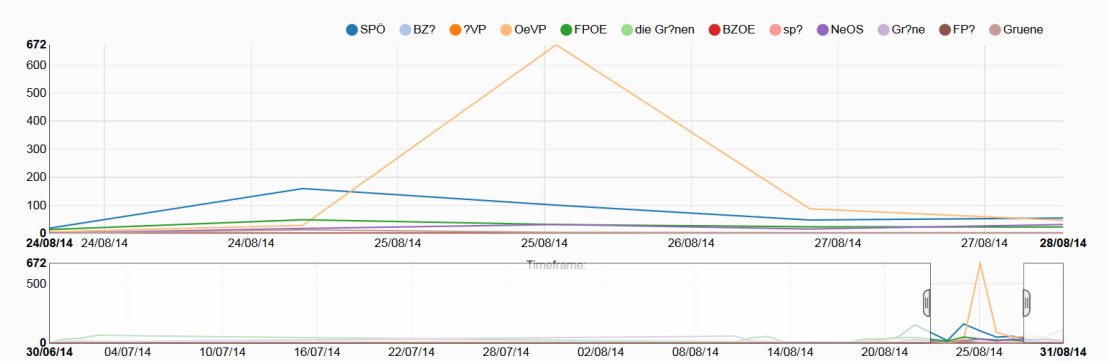


Figure 9. Further filtering along the time-axis

Next are two textual filters that introduce more flexibility, when zeroing in on the most relevant results in a Track.

- The "Entity mentions filter" functions similarly to the "Topics" option in the "Tracks"-Screen, accessing the TrendMiner ontology for existing entities that have been detected within that track.
- The "Plain text filter" takes into account the fact, that the TrendMiner ontology of political entities is not as dynamic as the political discourse and will often be unable to account for new topics, term, events and names that pop up in day-to-day discourse. It is therefore a plain text field where the user can enter any term they wish, to focus on results that they know to be in the data but that cannot be visualised or processed analytically since they have not yet been added to the ontology database.

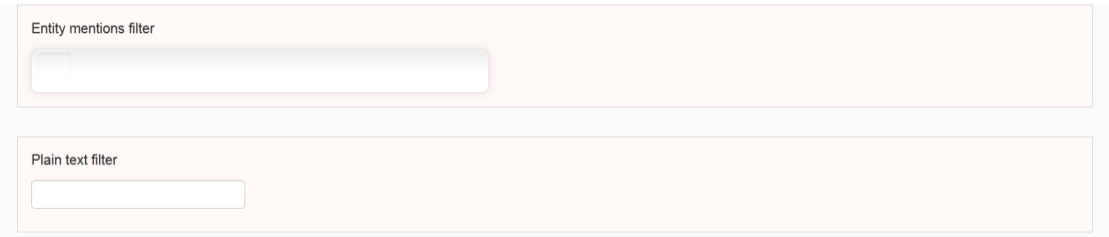


Figure 10. Further entity mentions filter and plain text filter

The final two elements of the Filter-page again serve dual functions. On the one hand they both display textual content of the track: Entities, like politicians, parties and authors of Tweets, Blogs or News articles. They also display topics and allow to filter tweets according to sentiment polarity.

Figure 11. Facetted browser for detected entities and tag-cloud

4 Future Planned Work

Future work will include the following activities:

- Additional annotated corpora will be created as part of the improvement of the political ontology, automated summary and sentiment analysis.
- As soon as the inclusion of additional media, such as RSS-Feeds is completed within the project, annotation of news texts will follow.
- Additional work on adapting existing resources, both ontologies and sentiment lexica will be carried out.
- The development of an indicator of emotional intensity will be pursued.
- Real-time, dynamic entity definition needed
- A possible spin-off application could be an analysis tool for textual data generated in the course of political research, like interview transcripts. Particularly commercial research firms, such as SORA, and their clients, stand to profit from a tool that saves on time and thus, costs.

5 Conclusion

We would like to summarize the work done in WP7 and in TrendMiner in general by listing some opportunities opened up by TrendMiner:

- Support PR professionals by providing overview & summarisation of big amounts of complex information
- Advance political science and consultancy through making fact-based expertise on relationship between public opinion and Twittersphere possible
- Enrich public discussion through making relationship between Social Media discourse and News cycles transparent to wider public.

- Multi-lingual real-time and historical analysis of streaming social media data TrendMiner aids the analysis of social media data both in both real-time and retrospective analysis of data sources like Twitter, Blogs and News Websites.
- Reliable identification of actors and topics in political discourse. Since political topics are constantly changing, with new terms and hashtags (in Twitter) being invented almost hourly, named entities like political figures, party politicians and public office holders, as well as political parties and public institutions or media serve as focal points to detect, visualise and structure topics in TrendMiner. Actors too change, but at a much slower rate, allowing the underlying ontologies to be updated by a human operator.
- Sentiment detection and analysis One aspect of analysis of social media in the field of politics is the ability to associate opinions and sentiments to relevant entities and topics mentioned in the data stream. For this both a lexical and grammatical approaches have been developed in TrendMiner: a lexical data base of polarity items has been developed, and some text analysis rules have been developed in order to relate such lexical items to the entitites that are mentioned in the data stream.
- Quantitative content analysis. TrendMiner aims to provide both "hard" and "soft" ways of analysing political discourse. It includes traditional "hard" numerical indicators that can give accurate and valid information about the public attention directed towards certain topics. This assists analysts in making sense of large datasets by breaking them down into numerical statistics, using quantitative content analysis methods. Indicators may give information on how frequently an entity (e.g. a person) or a topic (e.g. a hashtag) is mentioned, which other terms it co-occurs with and at what time. These statistics are useful on a very abstract level of analysis, e.g. for determining especially popular topics among a certain target group, or monitoring events live.
- Automated trend/topic detection and summaries. TrendMiner also supports "softer" ways of making sense of political discussions. Besides the traditonal word-clouds there are automatic topic detection, using spectral clustering methods. This generates clusters of terms (Hashtags, words, Names) that co-occur frequently within the same context, e.g. within Tweets, indicating real existing topics.
- Automated summaries to match existing documents (Tweets, sentences, whole artcles) to one or more ideal “average” tweets, generated from the selected subset (See D4.3.1).

Bibliography

Alm, C. O., D. Roth, and R. Sproat. “Emotions from Text: Machine Learning for Text-Based Emotion Prediction.” In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 579–86, 2005. http://dl.acm.org/citation.cfm?id=1220648.

Bethard, S., H. Yu, A. Thornton, V. Hatzivassiloglou, and D. Jurafsky. “Automatic Extraction of Opinion Propositions and Their Holders.” In *2004 AAAI Spring Symposium on Exploring Attitude and Affect in Text*, 2224, 2004.

Bethard, S., H. Yu, A. Thornton, V. Hatzivassiloglou, and D. Jurafsky. “Extracting Opinion Propositions and Opinion Holders Using Syntactic and Lexical Cues.” *Computing Attitude and Affect in Text: Theory and Applications*, 2006, 125–41.

Bhowmick, Plaban Kumar, Anupam Basu, and Pabitra Mitra. “Classifying Emotion in News Sentences: When Machine Classification Meets Human Classification.” *International Journal on Computer Science and Engineering* 2, no. 1 (2010): 98–108.

Ted Brader (2006): Campaigning for Hearts and Minds: How Emotional Appeals in Political Ads Work . Chicago: University of Chicago Press

Campbell, Angus, Phillip E. Converse, Warren E. Miller, and Stokes, Donald E. *The American Voter*. New York: Wiley, 1960.

Chmiel, Anna, and Janusz A. Hołyst. “Flow of Emotional Messages in Artificial Social Networks.” *arXiv:1004.4103v1*, 2010.

Conover, M. D., B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer. “Predicting the Political Alignment of Twitter Users.” In *Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom)*, 192–99, 2011. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6113114.

António R. Damásio, Hanna Damásio, Yves Christen: Neurobiology of Decision-Making, Berlin: Springer, 1996.

Demartini, G., S. Siersdorfer, S. Chelaru, and W. Nejdl. “Analyzing Political Trends in the Blogosphere.” In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewPDFInterstitial/2838/3244.

Downs, Anthony. *An Economic Theory of Democracy*. New York: Haroer & Row, 1957.

Ekman, P. (1972). Universals and Cultural Differences in Facial Expression of Emotion. In J. Cole ed. *Nebraska Symposium on Motivation*. Lincoln, Nebraska: University of Nebraska Press: 207-283.

Gayo-Avello, Daniel. “A Meta-Analysis of State-of-the-Art Electoral Prediction from Twitter Data.” *arXiv:1206.5851v1*, 2012. db/journals/corr/corr1206.html#abs-1206-5851.

Himelboim, Itai, Stephen McCreery, and Marc Smith. “Birds of a Feather Tweet Together: Integrating Network and Content Analyses to Examine Cross-Ideology Exposure on Twitter.” *Journal of Computer-Mediated Communication* 18, no. 2 (January 2013): 40–60. doi:10.1111/jcc4.12001.

Hofinger Christoph, and Gerlinde Manz-Christ. *How Neuroscience, Linguistics and Social Psychology Change the Political Profession*. Sydney: Prestige Books, 2011.

Inkpen, Diana, Fazel Keshtkar, and Diman Ghazi. “Analysis and Generation of Emotions in Text.” In *Selected Papers*. Cluj-Napoca, 2009.

Kim, S. M., and E. Hovy. “Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text.” In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, 1–8, 2006. http://dl.acm.org/citation.cfm?id=1654642.

Lampos, Vasileios. “On Voting Intentions Inference from Twitter Content: A Case Study on UK 2010 General Election.” *arXiv:1204.0423v7*, 2012. http://arxiv.org/pdf/1204.0423v7.pdf.

Lampos, Vasileios, Daniel Preotiuc-Pietro, and Trevor Cohn. “A User-Centric Model of Voting Intention from Social Media.” In *ACL (1)*, 993–1003, 2013. http://staffwww.dcs.shef.ac.uk/people/T.Cohn/pubs/lampos13bilinear.pdf.

Lazarsfeld, Paul, Bernard Berelson, and Hazel Gaudet. *The People´s Choice: How the Voter Makes up His Mind in a Presidential Campaign*. New York: Columbia University Press, 1944.

Liu, Bing. “Sentiment Analysis and Opinion Mining.” *Synthesis Lectures on Human Language Technologies* 5, no. 1 (2012): 1–167.

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)

Maynard, Diana, Kalina Bontcheva, and Dominic Rout. “Challenges in Developing Opinion Mining Tools for Social Media.” In *Workshop at LREC*. Istanbul, 2012. http://gate.ac.uk/sale/lrec2012/ugc-workshop/opinion-mining-extended.pdf.

Maynard, D., and A. Funk. “Automatic Detection of Political Opinions in Tweets.” In *The Semantic Web: ESWC 2011 Workshops*, 88–99, 2012. http://www.springerlink.com/index/F3H0J8N38010R730.pdf.

Metaxas, P. T., Mustafaraj, E., and Gayo-Avello, D. 2011. How (Not) To Predict Elections. In IEEE 3rd International Conference on Social Computing (SocialCom).

Morris Fiorina: Economic Retrospective Voting in American National Elections: A Micro-Analysis. American Journal of Political Science Vol 22 Nr. 2 426-443, 1978
Samuel Popkin: The reasoning voter. Communication and Persuasion in Presidential Campaigns. University of Chicago Press, 1991

O'Keefe, Tim, Tim O'Keefe James R. Curran, Peter Ashwell, and Peter Ashwell Irena Koprinska. "An Annotated Corpus of Quoted Opinions in News Articles" Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL) (2013). http://www.tokeefe.org/blog/wp-content/uploads/2013/08/acl13shortopinions.pdf.

Pang, Bo, and Lillian Lee. "Opinion Mining and Sentiment Analysis." *Foundations and Trends in Information Retrieval* 2, no. 1–2 (2008): 1–135.

Roberts, Kirk, Michael A. Roach, Joseph Johnson, Josh Guthrie, and Sanda M. Harabagiu. "EmpaTweet: Annotating and Detecting Emotions on Twitter." In *LREC*, 3806–13, 2012. http://lrec.elra.info/proceedings/lrec2012/pdf/201_Paper.pdf.

Mohammad, Saif. "Portable Features for Classifying Emotional Text," 587–91. Montréal, Canada, 2012.

Schreiber, D., & Fonzo, G. (2012). Throwing a big party? Neurocorrelates of membership in the major political parties. Neurocorrelates of Membership in the Major Political Parties.

Stieglitz, Stefan, and Linh Dang-Xuan. "Social Media and Political Communication: A Social Media Analytics Framework." *Social Network Analysis and Mining* 3, no. 4 (December 2013): 1277–91. doi:10.1007/s13278-012-0079-3.

Strapparava, C., and A. Valitutti. "WordNet-Affect: An Affective Extension of WordNet." In *Proceedings of LREC*, 4:1083–86, 2004. http://hnk.ffzg.hr/bibl/lrec2004/pdf/369.pdf.

Strapparava, C., and R. Mihalcea. "Learning to Identify Emotions in Text." In *Proceedings of the 2008 ACM Symposium on Applied Computing*, 1556–60, 2008. http://dl.acm.org/citation.cfm?id=1364052.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. 2010. Predicting elections with Twitter: What 140
characters reveal about political sentiment. In International AAAI Conference on Weblogs and Social Media. 178-185.

Valitutti, A., C. Strapparava, and O. Stock. "Developing Affective Lexical Resources." *PsychNology Journal* 2, no. 1 (2004): 61–83.
Wiegand, M., and D. Klakow. "The Role of Predicates in Opinion Holder Extraction," 2011. http://eprints.pascal-network.org/archive/00008849/.

Yasseri, Taha, and Jonathan Bright. “Can Electoral Popularity Be Predicted Using Socally Gener Ated Big Data?” *Information Technology* 56, no. 5 (2014): 246–5