Knowledge-Based Systems xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis

Sven Rill [a,b,*], Dirk Reinel [a], Jörg Scheidt [a], Roberto V. Zicari [b]

[a] Institute of Information Systems, University of Applied Sciences Hof, Alfons-Goppel-Platz 1, Hof, Germany
[b] Institute of Computer Science, Big Data Lab, Goethe University Frankfurt, Robert-Mayer-Str. 10, Frankfurt am Main, Germany

ARTICLE INFO

Article history:
Received 22 November 2013
Received in revised form 2 May 2014
Accepted 5 May 2014
Available online xxxx

Keywords:
Topic detection
Concept-level sentiment analysis
Big data
Twitter
Social data analysis

ABSTRACT

In this work, we present a system called *PoliTwi*, which was designed to detect emerging political topics (*Top Topics*) in Twitter sooner than other standard information channels. The recognized *Top Topics* are shared via different channels with the wider public. For the analysis, we have collected about 4,000,000 tweets before and during the parliamentary election 2013 in Germany, from April until September 2013. It is shown, that new topics appearing in Twitter can be detected right after their occurrence. Moreover, we have compared our results to Google Trends. We observed that the topics emerged earlier in Twitter than in Google Trends.

Finally, we show how these topics can be used to extend existing knowledge bases (web ontologies or semantic networks) which are required for concept-level sentiment analysis. For this, we utilized special Twitter hashtags, called *sentiment hashtags*, used by the German community during the parliamentary election.

1. Introduction

The impact of social networks like Facebook or Twitter has increased over the past few years, especially in the area of politics. In particular, Twitter provides a platform on which discussions on various topics can be detected sooner than other standard information channels. The following real world example illustrates this. During a press conference in the course of the visit of US President *Barack Obama* in Berlin, the German Federal Chancellor *Angela Merkel* said: "Das Internet ist für uns alle Neuland." – *'The Internet is new territory for all of us.'*. This statement was made on June 19, 2013 at 12:48 p.m. Already at 12:50 p.m., there was a first tweet with a hashtag "#neuland" – *'new territory'* on Twitter: "#neuland. Ich glaub Angela Merkel hat jetzt ein Meme an der Backe." – *'new territory. I think Angela Merkel is burdened by a meme now.'*. Minutes later, many similar tweets were written making '#neuland' a new *Top Topic* in Twitter. This example perfectly demonstrates how quickly news spread via Twitter. In other media that topic emerged at a later time. Major German online journals like *Süddeutsche*,[1] *ZEITOnline*[2] or *N24*[3] initially reported about the chancellor's statement not before 03:19 p.m.

In this paper, we describe a system designed to detect political topics emerging in Twitter. The main focus lies on a fast detection based on a few tweets at an early stage of a discussion. The data were collected before and during the parliamentary elections in 2013 in Germany. The system stayed in operation also after the election. Thus, data concerning political discussions have been collected from the beginning of April 2013. All topics extracted since the start of the project in April 2013 are available at our project website.[4]

Furthermore, we have extended our system by a sentiment analysis component to detect the polarity of topics marked by hashtags. For this, we use special Twitter hashtags, called *sentiment hashtags* (see Section 4), which people use to tag their opinion about politicians or parties. Our idea is to build up relation graphs for emerging political topics enriched with information like context and polarity. These graphs can later be used to extend an existing web ontology or a semantic network by a new dimension.

* Corresponding author at: Institute of Information Systems, University of Applied Sciences Hof, Alfons-Goppel-Platz 1, Hof, Germany. Tel.: +49 92814096312.
E-mail addresses: srill@iisys.de (S. Rill), dreinel@iisys.de (D. Reinel), jscheidt@iisys.de (J. Scheidt), zicari@informatik.uni-frankfurt.de (R.V. Zicari).

[1] http://www.sueddeutsche.de/politik/kritik-an-merkels-internet-aeusserung-neuland-aufschrei-im-spiesser-netz-1.1700710.
[2] http://www.zeit.de/digital/internet/2013-06/merkel-das-internet-ist-fuer-uns-alle-neuland.
[3] http://www.n24.de/n24/Nachrichten/Politik/d/3026394/merkel-erntet-spott-und-hohn.html.
[4] http://www.politwi.de, website only available in German.

http://dx.doi.org/10.1016/j.knosys.2014.05.008
0950-7051/© 2014 Published by Elsevier B.V.

This will contribute to improve concept-level sentiment analysis methods that use such knowledge bases.

2. Related work

Several publications about the behavior of Twitter users during election periods have been published. In [1–3], the researchers investigated the possibilities to predict election results on the basis of tweets. Gayo-Avello [4], on the other hand, shows that such a prediction is impossible. The goal of our project *PoliTwi* is not an election forecast but the detection of upcoming topics.

Data from the microblogging platform Twitter, i.e., tweets with a maximum length of 140 characters, is a special challenge for topic detection algorithms. Hong and Davison [5] investigated how standard topic models like *LDA* or *author-topic model* can be used for analyzing Twitter data. The *author-topic model* was introduced by Rosen-Zvi et al. [6]. Mathioudakis and Koudas [7] used an own algorithm named *QueueBurst* for the detection of topics and trends.

Some publications are based on the novel concept-level sentiment analysis. In [8], the authors discuss the evolution of opinion mining and sentiment analysis with a special focus on the existing approaches. Cambria et al. [9] introduce concept-level techniques that are based on statistical approaches. They present five different articles which cover varied approaches and solutions in this field. The first one discusses how sentiment classification can be adapted to specific domains [10]. In the second and third paper, an application in the area of customer reviews is described [11,12]. The forth [13] and the fifth [14] article cover the application of sentiment analysis for Spanish online videos and movie reviews, respectively.

Cambria [15] gives an overview about the field of concept-level sentiment analysis. He shows, how syntactics, semantics and pragmatics must overlap in order to build up algorithms and resources. Cambria and White [16] depict the recent developments in natural language processing research.

Algorithms for the automatic extraction of sentiments from textual data need text resources like polarity lexicons. These lexicons can be produced manually or derived from dictionaries or other text corpora. As there are at least several hundreds of commonly used opinion words in any language, a pure manual collection and rating of them is not feasible.

For the English language, commonly used resources are SentiWordNet (SWN) [17–19], Semantic Orientations of Words (SOW) [20], the Subjectivity Lexicon (SL) [21], SenticNet [22] and two lists of positive and negative opinion words [23]. SWN and SOW were generated using the WordNet® [24] lexical database. Lists of opinion values also exist for other languages [25–28].

In [29], the authors describe an approach to generate lists of opinion phrases including shifter or negation words. A German list was published in [30]. In [31], the authors propose a graph based model and its associated techniques to automatically acquire words' senses. In [32], the authors present two models that use interval type-2 fuzzy sets for representing the meaning of words that refer to emotions. They experimentally evaluate the use of the first model for translations and mappings between vocabularies.

The dictionary based approach has the disadvantage that it does not take into account that the polarities of many opinion words vary depending on their context. Cambria et al. [33] provide a resource, called SenticNet 3, for concept-level sentiment analysis.

3. Topic detection

The goal of the *PoliTwi* project is the identification of current and emerging political topics (*Top Topics*) for the parliamentary elections in 2013 in Germany. This information shall be shared with the wider public.

Hashtags are important concepts in Twitter. Usually, new topics are quickly marked by new hashtags, picked up and used in follow-up tweets. Because of the limitation of 140 characters per tweet, abbreviations or short artificial words are often used as hashtags. Examples in the context of the German parliamentary election 2013 were "#btw13" for "Bundestagswahl 2013" – *'German parliamentary election 2013'* or "#groko" for "Große Koalition" – *'Grand Coalition'*. In the *PoliTwi* project, hashtags are used as candidates for *Top Topics*.

A *Top Topic* is characterized by a significantly higher current appearance compared to a previous time period. The basic idea of our approach is to compare the current number of tweets with hashtag H $N(H,t_0)$ to the number of tweets of the previous period $\overline{N}(H,t)$ taking into account the standard deviation σ of the distribution. For the calculation of σ, we assume a Gaussian distribution of the number of tweets in the previous period. Here t_0 denotes the current time slot, e.g., the current hour or day, and t a reference time window in the past.

Our approach takes into account the temporal change in the number of tweets rather than their absolute number. To decide if a topic is classified as a *Top Topic*, for each hashtag occurring in the tweets a *Topic Value (tv)* is calculated. This value has a range between −1.0 and 1.0. A value greater than 0.0 means that the interest in a topic is increasing.

The number of tweets for a defined time period is mapped to a *Topic Value* (see Fig. 1) using the following formula.

$$tv(H,t_0) = \frac{N(H,t_0) - \overline{N}(H,t) - \sigma_{\overline{N}(H,t)}}{N(H,t_0) + \overline{N}(H,t) + \sigma_{\overline{N}(H,t)}} \quad (1)$$

Suitable parameter values have to be defined for the time slot t_0 and the reference time window t (see Section 6.3).

4. Twitter data

Our data set consists of tweets which were collected via the Twitter Streaming API.[5] Fig. 2 shows the total number of tweets per day for a defined period. We used political words and phrases, like party names, as search terms (see Section 6.1).

At the beginning of the project, about five months before the election day, we collected about 20,000 tweets per day. This value increased until the election day. We observed a larger number of tweets per day during the TV debate (September 1, 2013), on the state election day in Bavaria (September 15, 2013), and on the election day (September 22, 2013). On the election day we collected about 240,000 tweets. During the TV debate between the two German top candidates *Angela Merkel* and *Peer Steinbrück*, we measured the highest tweet rate with up to 1,200 tweets per minute. Due to technical issues relating to the Twitter Streaming API, we were not able to collect all relevant tweets between July 4 and July 18, 2013. Therefore this time period was excluded from the analysis.

In the context of the election, some special hashtags, called *sentiment hashtags*, were sometime used in German tweets. These hashtags contain an additional sign at the end of each tag which depicts the opinion of the author ('+' positive and '−' negative). *Sentiment hashtags* are usually used to express an opinion about a party or a politician, e.g., *'#CDU+'*. They were already used to detect the Twitter users' sentiments during the election in 2009. Sentiment hashtags were introduced by B. Unterberg and S. Lobo in 2009.[6]

[5] https://dev.twitter.com/docs/streaming-apis.
[6] http://www.twitterbarometer.de.

Q1 Please cite this article in press as: S. Rill et al., PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis, Knowl. Based Syst. (2014), http://dx.doi.org/10.1016/j.knosys.2014.05.008

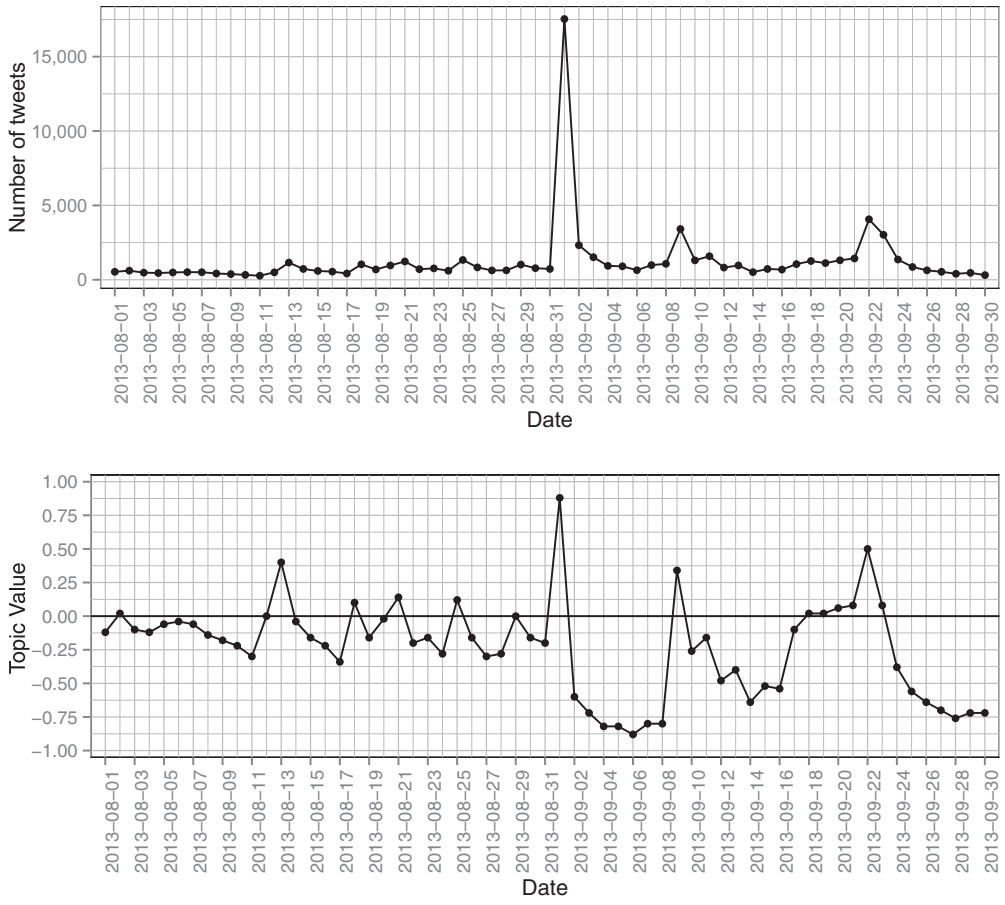

Fig. 1. Number of tweets and calculated *Topic Values* for hashtag '#merkel'.

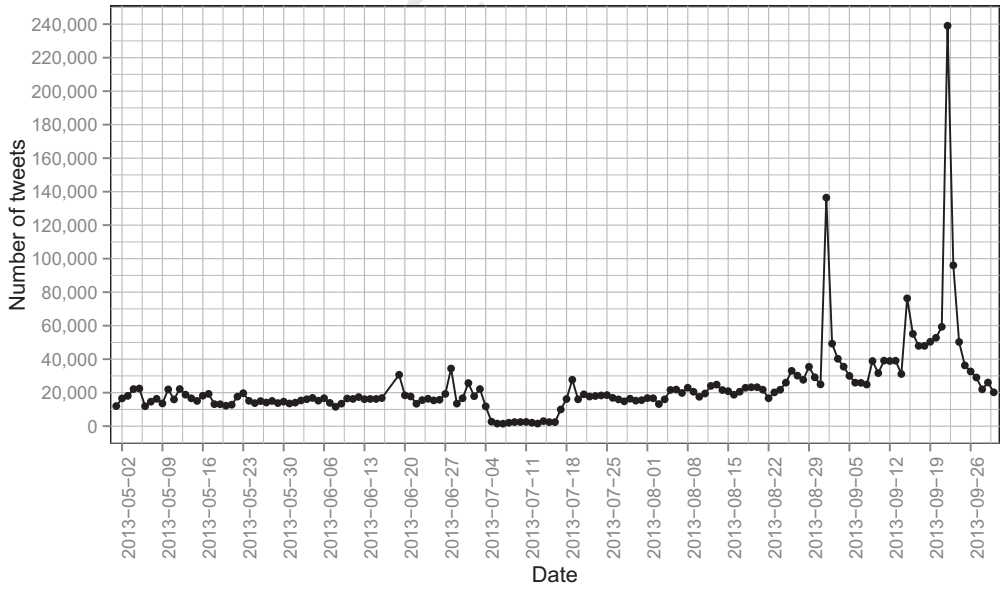

Fig. 2. Total number of tweets from May to early October 2013.

A summary of the collected Twitter data (May to October 2013):

- 4,000,000 tweets in total
- 2,500,000 user written tweets (see Section 6.2)
- 300,000 different users
- during TV debate: 100,000 tweets (up to 1200 tweets per minute)
- during election day: 240,000 tweets
- 97,000 extracted hashtags (used for topic detection)
- 100 different *sentiment hashtags*
- 171,000 tweets with *sentiment hashtags*

Q1 Please cite this article in press as: S. Rill et al., PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis, Knowl. Based Syst. (2014), http://dx.doi.org/10.1016/j.knosys.2014.05.008

5. Concept-level sentiment analysis

The basic idea of concept-level sentiment analysis is a deeper understanding of unstructured text by using a semantic approach. In order to achieve this goal, the usage of knowledge bases like SenticNet 3 or polarity lexicons like SWN, SOW and SL is essential. A major problem of this approach is the handling of novel words or topics, or words that are used in a new context, respectively. These words may get a different meaning in this new context, and thereby possibly obtain another polarity.

Our idea is to build a relation graph for the emerging political topics using the *sentiment hashtags* like *'#CDU+'* described above. This graph shall be enriched with a polarity probability. Fig. 3 shows an example for such a graph.

Originating from the hashtag '#neuland', which also is a *Top Topic*, one can see the connection to other words used in this context. We calculate the word frequency of all words included in all tweets with the hashtag '#neuland'. The relation graph contains the most frequently occurring words.

We would like to answer such question like: Which polarity bears an upcoming topic, e.g., the hashtag "#neuland", in this political context? How does the usage of this topic changes the polarity of a tweet?

To evaluate the polarity, we used *sentiment hashtags* and examined different time periods, concerning or not concerning the topic under investigation. We separated the period in which the new topic, e.g., "#neuland", came up and defined two reference time periods before and after the emerging of the new topic. Then, we built the graph (see Fig. 3) and analyzed, how the polarity of adjacent vertices, e.g., *CDU*, is changed by the presence of the new topic. To do so, we measured the polarity of *CDU* in the reference time periods and compared it to the time period under investigation. To find out whether the change of the polarity is significant, we used the binomial distribution to calculate the probability for the change being a statistical fluctuation.

We are now able to calculate the polarity of the emerging topics and to map the relations. Based on this information, an existing knowledge base can be expanded and concept-level sentiment analysis methods thus can base their analysis on those upcoming words and topics.

6. System implementation

The system consists of the modules *Data Selection*, *Preprocessing*, *Analysis* and *Presentation*. Fig. 4 gives an overview of the structure and the different modules.

6.1. Data selection

The tweets are collected using the Twitter Streaming API and stored in a raw database. For this purpose, a suitable crawler was implemented. It collects and stores tweets based on predefined search terms in a PostgreSQL database named *Crawler-DB*. The Twitter Streaming API includes several limitations in terms of the number of keywords as well as the maximum number of results.

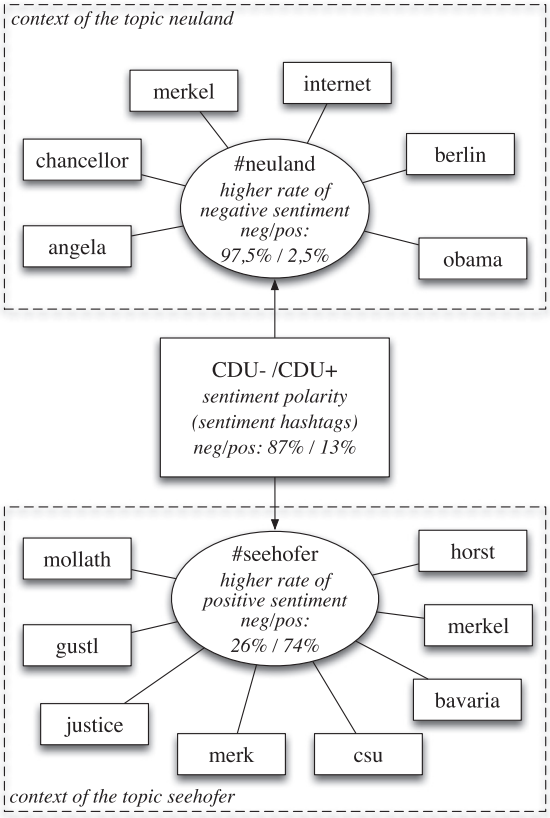

Fig. 3. Relation graph for two *Top Topics* in the context of *CDU* with their polarity.

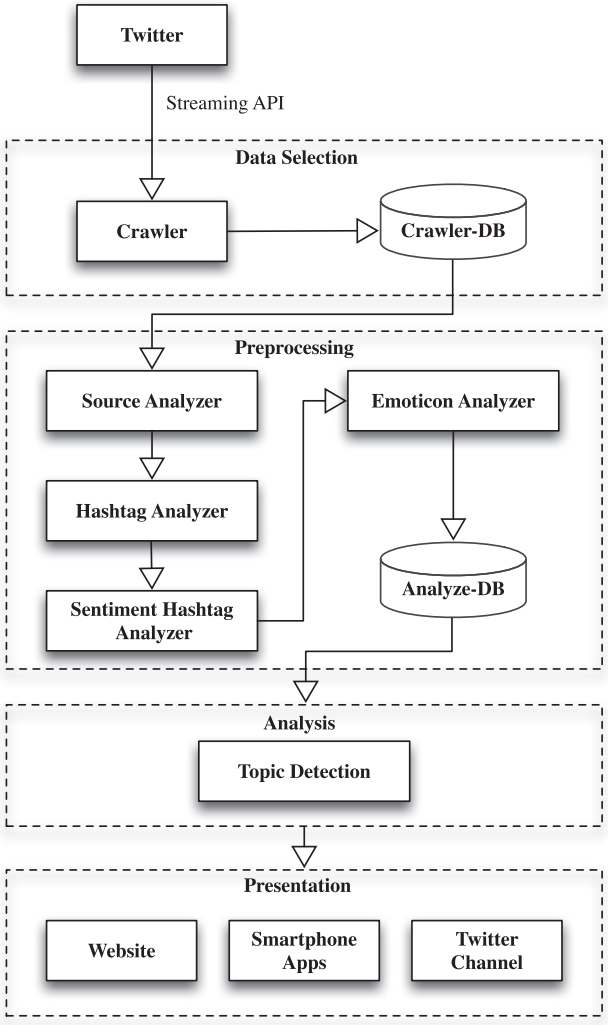

Fig. 4. System overview of *PoliTwi*.

Q1 Please cite this article in press as: S. Rill et al., PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis, Knowl. Based Syst. (2014), http://dx.doi.org/10.1016/j.knosys.2014.05.008

Table 1
Excerpt of search terms for using the Twitter Streaming API.

Party	Synonyms	Politicians	Synonyms
CDU	Christlich Demokratische Union, Union, Unionspartei, Christdemokraten	Angela Merkel	Angela, Merkel, Angie, Kanzlerin, Bundeskanzlerin, Frau Merkel
SPD	Sozialdemokratische Partei Deutschlands, Sozialdemokraten	Peer Steinbrück	Steinbrück
Grüne	Bündnis 90/Die Grünen, Grünen, Gruene, Gruenen, Die Grünen, Bündnis 90, Bündnis90	Jürgen Trittin, Katrin Göring-Eckardt	Trittin, Göring-Eckardt

During the project, these limitations were not reached, thus having no effect on the results.

Our search terms include nine major German parties (*CDU*, *CSU*, *SPD*, *Linke*, *Die Grünen*, *FDP*, *AfD*, *Piraten* and *NPD*), the names of the top candidates (e.g., *Angela Merkel* and *Peer Steinbrück*) as well as special terms related to the election (e.g., *btw13*). In addition, common synonyms for the politicians and parties were added (e.g., "Bundeskanzlerin" – *'Federal Chancellor'* or "Kanzlerin" – *'Chancellor'* for *Angela Merkel*). This resulted in a total number of 87 search terms. These terms were not changed during the project. Table 1 shows examples of search terms plus some of the synonyms used for the parties *CDU*, *SPD*, and *Die Grünen*.

With these search terms, some tweets without a political context have been collected. This occurs, whenever a term is used with polysemous meanings. For example, the word "grün" – *'green'* is used for the party called "Die Grünen" - *'The Greens'* but also as the color green in, e.g., "grüner Tee" – *'green tea'*.

For each tweet, a set of meta information is available from the Twitter Streaming API, e.g., the users' language or location. This meta information is also stored in the database and can be used for further analysis or extensions of the system.

6.2. Preprocessing

In this module, various preprocessing steps are performed.

The most promising candidates for detecting new topics in Twitter are user written tweets. In the best case, an eyewitness reports a new topic via Twitter and other users retweet or commend this topic. There are automatically generated tweets that we want to exclude in the *Source Analyzer*, as these tweets often just refer to an online article. An example for such a tweet is 'Die #GroKo will finanzielle Hilfen für #Windkraft radikal kürzen – für viele würde das Verluste bedeuten http://spon.de/ad4Qq (red)" – *'The grand coalition wants to radically cut down the financial support for wind power – for many this would mean losses http://spon.de/ad4Qq (red)'*.

For each tweet, the information about the source (*Twitter app*) is provided by the Twitter Streaming API. Twitter apps are, for example, the Twitter website, clients for various mobile platforms (like iOS, Android or Blackberry) or the Content Management Systems of news platforms, each having a unique identifier. For the identification of the Twitter apps used by real Twitter users, a whitelist has been defined.

To define the whitelist, a sample of 1,000,000 tweets was evaluated. We considered only Twitter apps that generate more than 500 tweets. We ended up with 75 Twitter apps and a share of 95.12% of the 1,000,000 tweets. Excluding all Twitter apps used for the automated generation of tweets leads to 26 remaining apps and a share of 75.01% of the 1,000,000 tweets being used for the topic detection. An example of an app used by real Twitter users is 'Twitter for iPhone'.

The *Hashtag Analyzer* extracts all hashtags contained in the tweets. These hashtags are the basis for the topic detection.

The *Sentiment Hashtag Analyzer* marks all *sentiment hashtags* (see Section 4) contained in the tweets. In the data set, about 100 different *sentiment hashtags* in 171,000 tweets can be found.

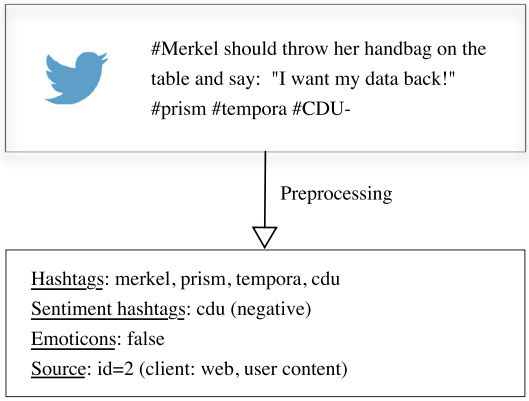

Fig. 5. Preprocessing results for an example tweet.

The *Emoticon Analyzer* extracts information about emoticons inside of the tweets, often seen as markers for the presence of irony in such messages. This information is not yet exploited for this study. It will be used in a later extension of the system.

The results of the preprocessing, which are stored in the so called *Analyze-DB*, are the basis for the topic detection later on. Fig. 5 shows a tweet and the results of the preprocessing.

6.3. Analysis

In the module *Analysis*, the topic detection is carried out as described in Section 3.

Here, only those hashtags contained in the tweets are candidates for *Top Topics*.

For the presentation of the results of the topic detection, only *Top Topics* with a *Topic Value* (see Section 3) greater than 0.0 are selected. In addition, the scale of the *Topic Values* is transformed to a range between 0 and 100. We call the resulting measure *Topic Points*.

Top Topics are presented for both, the hour preceding the current hour and the day preceding the current day. Thus, the time slots were fixed to one hour and one day. The calculation of $\overline{N}(H,t)$, necessary for the calculation of the *Topic Value* according to Eq. (1) given in Section 3, is done as follows:

$$\overline{N}(H,t) = \frac{\sum_{i=1}^{n_t} N(H, t_0 - offset - i)}{n_t} \quad (2)$$

Here, H stands for the Twitter hashtag under consideration, t for the time slot (day or hour), $\overline{N}(H,t)$ denotes the mean value of the number of tweets for the hashtag H and the reference time slot t. The number of days or hours in the reference time window is denoted as n_t, respectively. t_0 denotes the current day or hour. The *offset* allows to shift the reference time window. It is set to zero for the calculation of the *Top Topics* on a per day basis. The effect is that a topic can remain a *Top Topic* for several hours, even if the number of tweets for this topic slightly decreases. Finally, $N(H,t)$ is the number of tweets for a given hashtag and a given time period.

Q1 Please cite this article in press as: S. Rill et al., PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis, Knowl. Based Syst. (2014), http://dx.doi.org/10.1016/j.knosys.2014.05.008

Retweets are considered as regular tweets. Retweeting is one of the main mechanisms in Twitter, designed to help users in quickly distributing new information. Hence, it is appropriate to take these retweets into consideration when deriving the *Top Topics*.

6.4. Presentation

The purpose of the module *Presentation* is the presentation of the current *Top Topics* to the wider public. Several channels are used, i.e., a Twitter channel, a website as well as smartphone apps.

6.4.1. Twitter channel

The system automatically generates tweets announcing the actual *Top Topics* every full hour and daily at midnight. For this, a Twitter channel[7] is used. We communicate as many *Top Topics* as fit into a single tweet. They are sorted according to the *Topic Value*. The Twitter channel turned out to be the most effective way for distributing the *Top Topics*. Meanwhile, over 1000 users follow this channel. About 200 of them are political parties, politicians and journalists. These users have more than 1,500,000 followers in total. This leads to the effect that, whenever one of our followers retweets one of the *PoliTwi* tweets, this tweet reaches many other users.

6.4.2. Website

As the Twitter channel only allows the distribution of short lists of *Top Topics*, much more information is available on the website:

- Total number of tweets: Shows the total number of all analyzed tweets.
- Number of tweets for certain hashtags: The number of analyzed tweets for the two top candidates (*#Merkel* and *#Steinbrück*) are displayed.
- Topics for the preceding hour/day: The *Top Topics*, as described in Section 6.3, are displayed on a hourly ("Top-Themen der vorherigen Stunde" – *'Top Topics of the preceding hour'*) and daily ("Top-Themen des Vortages" – *'Top Topics of the preceding day'*) basis. Per *Top Topic*, the *Topic Points* as well as the absolute number of tweets are shown. The *Top Topics* are listed according to the *Topic Points* in descending order.
- On the web page "Selber forschen" – *'Do your own research'*, all former *Top Topics* can be accessed. The presentation is realized with DataTable (designed and created by Allan Jardine[8]). A search for *Top Topics* is possible as well as a sorting according to the several columns displayed.
- The depiction of the chronological sequence of the *Top Topics* is realized with the help of the Google Chart tools.[9] It is possible to trace the number of tweets for a given *Top Topic* within the given time window.

Currently, the website is only available in German. Figs. 6 and 7 are showing translated versions.

6.4.3. Smartphone apps

In addition to the website, also native smartphone apps were developed for the two most important platforms (iOS[10] and Android[11]).

[7] https://twitter.com/politwi.
[8] http://www.sprymedia.co.uk/.
[9] https://developers.google.com/chart/.
[10] https://itunes.apple.com/de/app/politwi/id688149225.
[11] https://play.google.com/store/apps/details?id=de.iisys.ais.politwi.androidapp.

7. Experiments and results

We describe some experiments on topic detection and the concept-level sentiment analysis and show first results.

7.1. Topic detection

For our experiments, we chose Google Trends[12] as a reference data set. In Google Trends, the fraction of the occurrence of a search term in relation to all search terms is displayed. Based on this, Google Trends calculates a value in a range between 0.0 and 1.0 for each search term on a daily basis. We compared certain *Top Topics* with the data from Google Trends. To ease the comparison, we re-scaled the *Topic Values* to a scale with values between 0.0 and 1.0. We calculated the Pearson product-moment correlation coefficient for each *Top Topic*, assuming a linear relation between the Google Trends data and our *adjusted Topic Values*.

7.1.1. Correlation for Topic "Neuland" – 'new territory'

The *Top Topic* 'Neuland' was brought up by *Angela Merkel* on June 19 (see Section 1). Fig. 8 depicts the *adjusted Topic Values* in comparison to the Google Trends values. The correlation coefficient is 0.68 with a p-value of 3.1×10^{-5}.

7.1.2. Correlation for topic 'Snowden'

The *Top Topic* 'Snowden' emerged on June 9 when an interview with the *Guardian* was published. Fig. 9 depicts the *Topic Values* compared to the Google Trends values. The correlation coefficient is 0.67 with a p-value of 5.2×10^{-5}.

7.1.3. Time shift between Twitter and Google

The shape of the graph in Fig. 8 led to the hypothesis that the *Topic Values* precede the Google Trends values. To verify this hypothesis, we calculated the correlation coefficient ($corr_1$) and the coefficient of determination (R_1^2) between the *Topic Values* of the current day in comparison to the Google Trends values of the day after. Afterwards, we compared the results to the values obtained in Sections 7.1.1 and 7.1.2, denoted as $corr_0$. Table 2 summarizes the results.

As expected, the correlation coefficients increased significantly when comparing the *PoliTwi Topic Values* with the Google Trends values of the day after, indicating that the emergence of a topic on Twitter precedes the Google Trends for this topic.

7.2. Concept-level sentiment analysis

We also used three *Top Topics* to perform an experiment with regard to concept-level sentiment analysis. We want so examine, whether a *Top Topic* changes the polarity of the adjacent vertex *CDU* in the graph shown in Fig. 3.

Therefore, we used the *sentiment hashtags* (see Section 4). As described in Section 5, we divided the data into three 30 days long time periods. The test period covers the month in which the topic was detected. The other two time periods are the months before (reference period 1) and after (reference period 2) the test period.

We counted the number of positive and negative tweets containing the hashtags *'#CDU+'* and *'#CDU−'* in the reference periods 1 and 2 and compared these numbers to the number of tweets in the test period having also the hashtags '#neuland', "#hochwasser" – *'#flood'* and '#seehofer' (German politician). Table 3 summarizes the numbers of positive and negative tweets in the two reference time periods. Table 4 summarizes the numbers of positive and negative tweets for three *Top Topics*.

[12] http://www.google.com/trends/.

Q1 Please cite this article in press as: S. Rill et al., PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis, Knowl. Based Syst. (2014), http://dx.doi.org/10.1016/j.knosys.2014.05.008

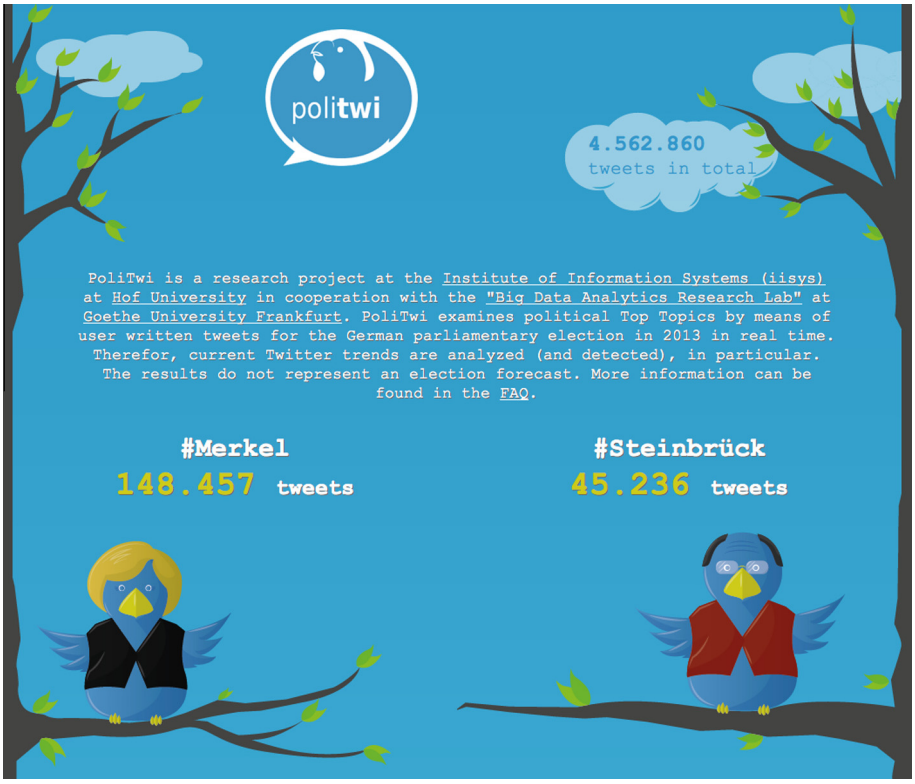

Fig. 6. Main page of the *PoliTwi* website with general information. (translated).

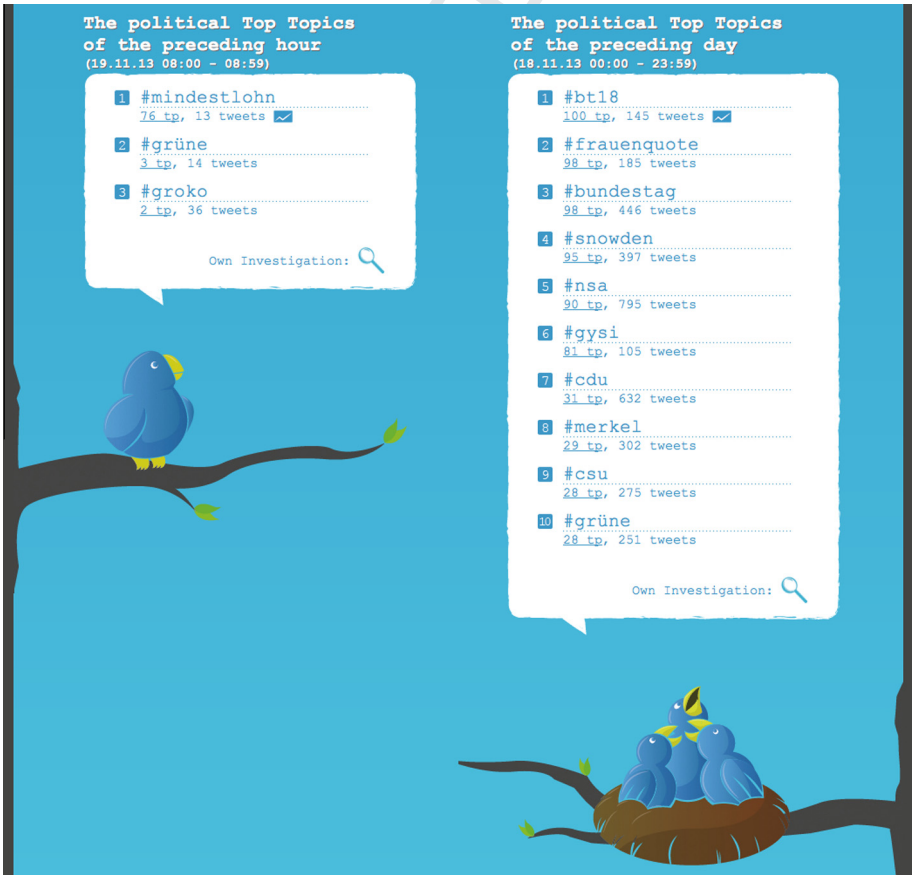

Fig. 7. Presentation of the *Top Topics* at the *PoliTwi* website. (translated).

Q1 Please cite this article in press as: S. Rill et al., PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis, Knowl. Based Syst. (2014), http://dx.doi.org/10.1016/j.knosys.2014.05.008

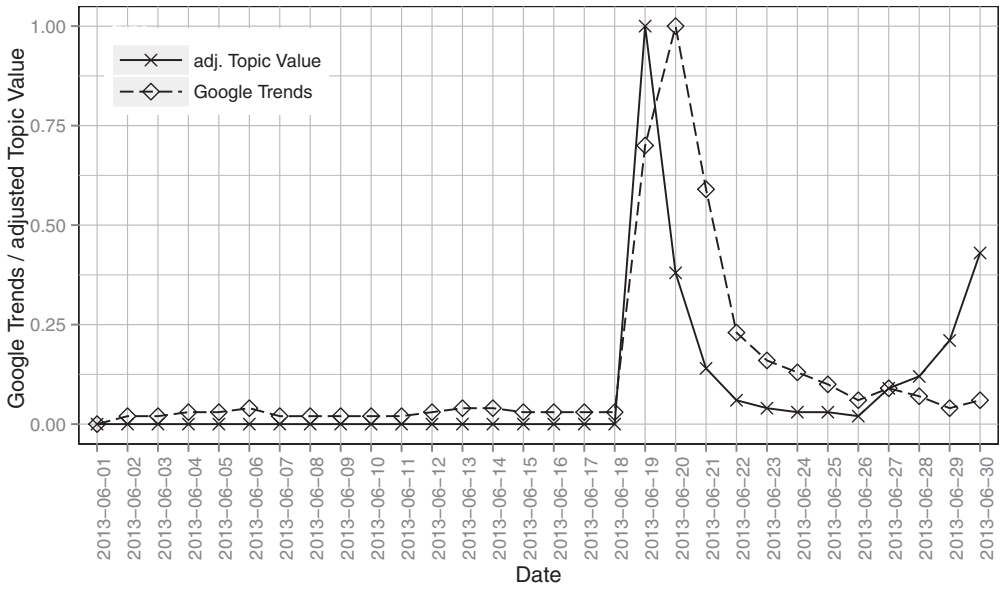

Fig. 8. Comparison between *PoliTwi* and Google Trends for the *Top Topic* "Neuland" – *'new territory'*.

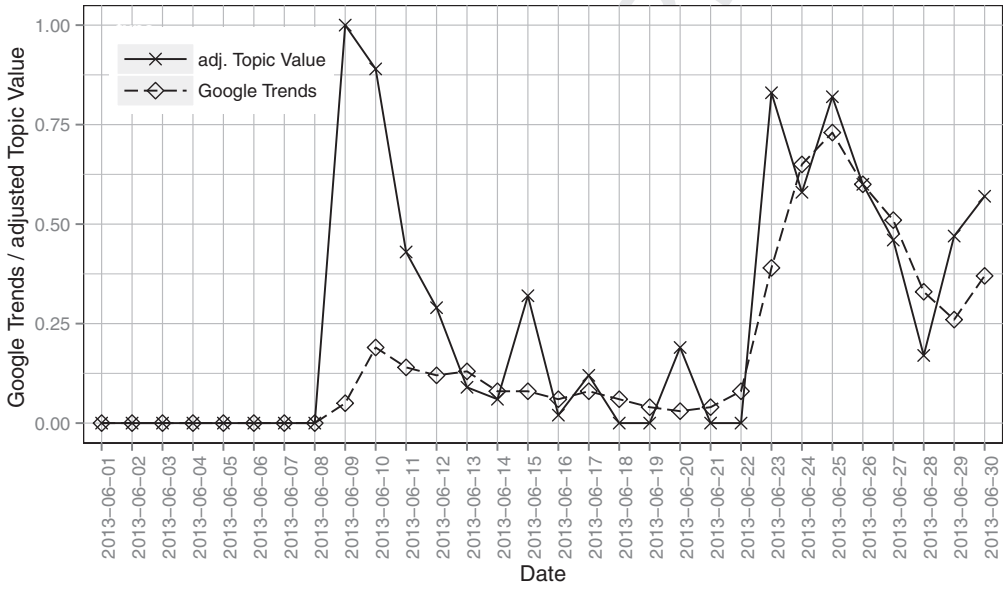

Fig. 9. Comparison between *PoliTwi* and Google Trends for the *Top Topic* 'Snowden'.

Table 2
Correlation results between *PoliTwi* and Google Trends.

Topic	$corr_0$	R_0^2	p-Value$_0$	$corr_1$	R_1^2	p-Value$_1$
Neuland – *new territory*	0.68	0.47	3.1e−05	0.81	0.65	6.7e−08
Snowden	0.67	0.45	5.2e−05	0.68	0.47	3.0e−05

Table 3
Number of tweets with sentiment hashtags *'#CDU+'* and *'#CDU−'* in the two reference periods.

Sentiment hashtag	Reference period 1 Count (Fraction)	Reference period 2 Count (Fraction)
CDU+	98 (0.139)	708 (0.131)
CDU−	607 (0.861)	4,693 (0.869)
Sum	705	5401

Table 4
Number of tweets with sentiment hashtags *'#CDU+'* and *'#CDU−'* and three *Top Topics*.

Sentiment hashtag	Test period		
	Topic Neuland Count (Fraction)	Topic Hochwasser Count (Fraction)	Topic Seehofer Count (Fraction)
CDU+	5 (0.025)	11 (0.118)	31 (0.738)
CDU−	194 (0.975)	82 (0.882)	11 (0.262)
Sum	199	93	42

To check whether the small number of positive or negative tweets for a *Top Topic* in the time period, in which this topic was detected, could be a statistical fluctuation, the probability of a number of k or less positive/negative tweets in a sample of N tweets was calculated using the binomial distribution (see Eq. (3)).

Q1 Please cite this article in press as: S. Rill et al., PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis, Knowl. Based Syst. (2014), http://dx.doi.org/10.1016/j.knosys.2014.05.008

Table 5
Results of the calculation using the binomial distribution for the three *Top Topics*.

	Topic Neuland	Topic Hochwasser	Topic Seehofer
$P_{acc,ref1}$	3.72×10^{-8}	0.346	2.37×10^{-18}
$P_{acc,ref2}$	1.67×10^{-7}	0.431	4.16×10^{-19}

$$P_{acc,topic,\leqslant k} = P(N_{pos/neg} \leqslant k) = \sum_{i=0}^{k} \binom{n}{i} \cdot P_{pos/neg}^{i} \cdot P_{neg/pos}^{N-i} \quad (3)$$

We got the following results (see Table 5) for three *Top Topics* and the two reference periods (*ref1* and *ref2*).

These numbers indicated that a statistical fluctuation could be excluded for the *Top Topics* 'Neuland' and 'Seehofer'. In the context of the *Top Topic* 'Neuland', the adjacent vertex *CDU* has a more negative polarity. In the context of the *Top Topic* 'Seehofer', it has a more positive polarity. In the context of the *Top Topic* 'Hochwasser', it has no significant impact of the polarity.

With this approach, we can calculate how the polarity of a known concept changes in the context of new topics.

8. Future work

Planned extensions of the project cover the following issues:

- Inclusion of sentiment analysis to investigate whether controversial topics arise faster than others.
- Exploitation of additional meta information stored with each tweet, e.g., the geo-information, to examine the spatial distribution of the tweets.
- Investigation of the possibility to derive a context from jointly occurring political topics.

Since the concept-level sentiment analysis part only is a concept, which was successfully examined with some experiments by now, we plan a technical implementation and further tests. In the course of this, especially the described approach to extend an existing knowledge base is to be tested.

Furthermore, the project can be the basis for additional services, e.g., notifications or individually configurable analyses, offered to interested users.

9. Conclusion

In this work, we have demonstrated how to quickly detect new emerging topics using data from Twitter.

We analyzed tweets before and after the parliamentary election in 2013 in Germany. The data collection from Twitter started five months before the election and is still ongoing.

We present our results, current *Top Topics* as well as all former *Top Topics*, to the general public via several channels. Especially the Twitter channel is widely used by journalists and people interested in politics.

The data collected allows a broad variety of further investigations.

Furthermore, we showed that it is possible to calculate the polarity of emerging topics. We described an idea how information like relations between topics and the polarity can be used to extend existing knowledge bases to improve concept-level sentiment analysis methods.

Acknowledgments

The authors would like to thank all members of the Institute of Information Systems (iisys) at the University of Applied Sciences Hof and also the members of the Big Data Lab at the Goethe University Frankfurt for many helpful discussions. We also thank Richard Göbel for his contribution as to the foundation of the institute. The Institute of Information Systems is supported by the Foundation of Upper Franconia and by the State of Bavaria.

References

[1] M. Skoric, N. Poor, P. Achananuparp, E.-P. Lim, J. Jiang, Tweets and votes: a study of the 2011 Singapore general election, in: Proceedings of the 2012 45th Hawaii International Conference on System Sciences, HICSS '12, IEEE Computer Society, 2012, pp. 2583–2591.
[2] A. Tumasjan, T. Sprenger, P. Sandner, I. Welpe, Predicting elections with twitter: what 140 characters reveal about political sentiment, in: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 2010, pp. 178–185.
[3] A. Jungherr, Tweets and votes a special relationship: the 2009 federal election in Germany, in: Proceedings of the 2Nd Workshop on Politics, Elections and Data, PLEAD '13, ACM, 2013, pp. 5–14.
[4] D. Gayo-Avello, No, you cannot predict elections with twitter, IEEE Internet Comput. 16 (6) (2012) 91–94.
[5] L. Hong, B.D. Davison, Empirical study of topic modeling in twitter, in: Proceedings of the First Workshop on Social Media Analytics, SOMA'10, 2010, pp. 80–88.
[6] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, M. Steyvers, Learning author-topic models from text corpora, ACM Trans. Inf. Syst. 28 (1) (2010) 4:1–4:38.
[7] M. Mathioudakis, N. Koudas, Twittermonitor: trend detection over the twitter stream, in: Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, SIGMOD'10, 2010, pp. 1155–1158.
[8] E. Cambria, B. Schuller, Y. Xia, C. Havasi, New avenues in opinion mining and sentiment analysis, IEEE Intell. Syst. 28 (2) (2013) 15–21.
[9] E. Cambria, B. Schuller, B. Liu, H. Wang, C. Havasi, Statistical approaches to concept-level sentiment analysis, IEEE Intell. Syst. 28 (3) (2013) 6–9.
[10] R. Xia, C. Zong, X. Hu, E. Cambria, Feature ensemble plus sample selection: domain adaptation for sentiment classification, IEEE Intell. Syst. 28 (3) (2013) 10–18.
[11] L. Garcia-Moya, H. Anaya-Sanchez, R. Berlanga-Llavori, Retrieving product features and opinions from customer reviews, IEEE Intell. Syst. 28 (3) (2013) 19–27.
[12] G. Di Fabbrizio, A. Aker, R. Gaizauskas, Summarizing online reviews using aspect rating distributions and language modeling, IEEE Intell. Syst. 28 (3) (2013) 28–37.
[13] V. Rosas, R. Mihalcea, L. Morency, Multimodal sentiment analysis of spanish online videos, IEEE Intell. Syst. 28 (3) (2013) 38–45.
[14] M. Wollmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, L.-P. Morency, Youtube movie reviews: Sentiment analysis in an audio-visual context, IEEE Intell. Syst. 28 (3) (2013) 46–53.
[15] E. Cambria, An introduction to concept-level sentiment analysis, in: Advances in Soft Computing and Its Applications, Lecture Notes in Computer Science, vol. 8266, Springer, Berlin Heidelberg, 2013, pp. 478–483.
[16] E. Cambria, B. White, Jumping nlp curves: a review of natural language processing research [review article], IEEE Comput. Intell. Mag. 9 (2) (2014) 48–57.
[17] S. Baccianella, A. Esuli, F. Sebastiani, Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, in: Proceedings of the 7th International Conference on Language Resources and Evaluation, 2010, pp. 2200–2204.
[18] A. Esuli, F. Sebastiani, Determining term subjectivity and term orientation for opinion mining, in: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, 2006, pp. 193–200.
[19] A. Esuli, F. Sebastiani, SentiWordNet: A publicly available lexical resource for opinion mining, in: Proceedings of the 5th International Conference on Language Resources and Evaluation, 2006, pp. 417–422.
[20] H. Takamura, T. Inui, M. Okumura, Extracting semantic orientations of words using spin model, in: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, 2005, pp. 133–140.
[21] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis, in: Proceedings of the Human Language Technology Conference, 2005, pp. 347–354.
[22] E. Cambria, R. Speer, C. Havasi, A. Hussain, Senticnet: A publicly available semantic resource for opinion mining., in: AAAI Fall Symposium: Commonsense Knowledge, vol. FS-10-02 of AAAI Technical Report, AAAI, 2010.
[23] B. Liu, M. Hu, J. Cheng, Opinion observer: Analyzing and comparing opinions on the web, in: Proceedings of the 14th International World Wide Web Conference, 2005, pp. 342–351.
[24] G.A. Miller, Wordnet: a lexical database for english, Commun. ACM 38 (1995) 39–41.
[25] J. Brooke, M. Tofiloski, M. Taboada, Cross-linguistic sentiment analysis: From english to spanish, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing, 2009, pp. 50–54.
[26] S. Clematide, M. Klenner, Evaluation and extension of a polarity lexicon for german, in: Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, 2010, pp. 7–13.

[27] R. Remus, U. Quasthoff, G. Heyer, Sentiws – a publicly available german-language resource for sentiment analysis, in: Proceedings of the 7th International Conference on Language Resources and Evaluation, 2010, pp. 1168–1171.
[28] U. Waltinger, Germanpolarityclues: A lexical resource for german sentiment analysis, in: Proceedings of the 7th International Conference on Language Resources and Evaluation, 2010, pp. 1638–1642.
[29] S. Rill, J. Drescher, D. Reinel, J. Scheidt, O. Schütz, F. Wogenstein, D. Simon, A Generic Approach to Generate Opinion Lists of Phrases for Opinion Mining Applications, in: Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM-12), 2012.
[30] S. Rill, S. Adolph, J. Drescher, D. Reinel, J. Scheidt, O. Schütz, F. Wogenstein, R. Zicari, N. Korfiatis, A phrase-based opinion list for the german language, in: Proceedings of the 1st Workshop on Practice and Theory of Opinion Mining and Sentiment Analysis (PATHOS), 2012, pp. 305–313.
[31] A. Jain, D.K. Lobiyal, A new method for updating word senses in hindi wordnet, in: 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014, pp. 666–671.
[32] A. Kazemzadeh, S. Lee, S. Narayanan, Fuzzy logic models for the meaning of emotion words, IEEE Comput. Intell. Mag. 8 (2) (2013) 34–49.
[33] E. Cambria, D. Olsher, D. Rajagopal, Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis, in: AAAI, 2014.

Q1 Please cite this article in press as: S. Rill et al., PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis, Knowl. Based Syst. (2014), http://dx.doi.org/10.1016/j.knosys.2014.05.008