Received 29 June 2016
Revised 10 July 2017
Accepted 11 July 2017

Using Twitter sentiment and emotions analysis of Google Trends for decisions making

Ernesto D'Avanzo
*Department of Political, Social and Communication Sciences,
University of Salerno, Salerno, Italy and
Institute of High Performance Computing and Networking (ICAR),
National Research Council of Italy (CNR), Palermo, Italy*
Giovanni Pilato
*Institute of High Performance Computing and Networking (ICAR),
National Research Council of Italy (CNR), Palermo, Italy, and*
Miltiadis Lytras
University of Patras, Patras, Greece

Abstract

Purpose – An ever-growing body of knowledge demonstrates the correlation among real-world phenomena and search query data issued on Google, as showed in the literature survey introduced in the following. The purpose of this paper is to introduce a pipeline, implemented as a web service, which, starting with recent Google Trends, allows a decision maker to monitor Twitter's sentiment regarding these trends, enabling users to choose geographic areas for their monitors. In addition to the positive/negative sentiments about Google Trends, the pipeline offers the ability to view, on the same dashboard, the emotions that Google Trends triggers in the Twitter population. Such a set of tools, allows, as a whole, monitoring real-time on Twitter the feelings about Google Trends that would otherwise only fall into search statistics, even if useful. As a whole, the pipeline has no claim of prediction over the trends it tracks. Instead, it aims to provide a user with guidance about Google Trends, which, as the scientific literature demonstrates, is related to many real-world phenomena (e.g. epidemiology, economy, political science).

Design/methodology/approach – The proposed experimental framework allows the integration of Google search query data and Twitter social data. As new trends emerge in Google searches, the pipeline interrogates Twitter to track, also geographically, the feelings and emotions of Twitter users about new trends. The core of the pipeline is represented by a sentiment analysis framework that make use of a Bayesian machine learning device exploiting deep natural language processing modules to assign emotions and sentiment orientations to a collection of tweets geolocalized on the microblogging platform. The pipeline is accessible as a web service for any user authorized with credentials.

Findings – The employment of the pipeline for three different monitoring task (i.e. consumer electronics, healthcare, and politics) shows the plausibility of the proposed approach in order to measure social media sentiments and emotions concerning the trends emerged on Google searches.

Originality/value – The proposed approach aims to bridge the gap among Google search query data and sentiments that emerge on Twitter about these trends.

Keywords Bayesian analysis, Emotion analysis, Social behaviour forecasting, Social sensing, Social sentiment analysis, Social trends

Paper type Research paper

1. Toward the near future: social trend-based decisions

Social media are more and more employed as indicators of public opinions of the real-world phenomena (Zhang *et al.*, 2014; Gao *et al.*, 2014; Xia *et al.*, 2014), from epidemiology, that tries to predict the crash of pandemic diseases (Ginsberg *et al.*, 2009), to economy, interested in

This work has been partially supported by the MIUR PON01 01687 – SINTESYS Research Project. The authors kindly thank graduate and undergraduate students for their support.

Program
Vol. 51 No. 3, 2017
pp. 322-350
© Emerald Publishing Limited
0033-0337
DOI 10.1108/PROG-02-2016-0015

how correlated are job-related queries with the rise and fall of unemployment rate. Furthermore, political science uses search query data to recognize patterns about the amount of political contributions raised by candidates. Other such examples can be found in the literature.

These ongoing trends exploit social search behaviors (Lagre *et al.*, 2015) and sentiments (Liu, 2005) as cues for consumers and sellers' decisions (Dhar and Simonson, 2003)[1]. In fact, thanks to the availability of the aggregated frequency of search queries[2] from some search engines services, such as those provided by Google Trends[3], these new kinds of search services, also known as query of search queries[4], become their own new social media the content of which, just trends of search, is created by users.

Actually, thanks to Google Trends, they have been demonstrated several examples of how the search volume for keywords coincides with as many patterns, showing how these kinds of correlation hold for many local phenomena. Think of the case of the keyword Summer camp, as reported by Webb (2009), whose number of searches seem to increase when the end of the school year is approaching, probably because of parents Summer recreation plans for their kids. The same seems to happen for Internal Revenue Service keyword, on April 15, when the tax deadline in USA is approaching.

But the story does not end here; in fact, it is just begun. For instance, it seems that media providers, more and more, look at Google Trends in order to determine the hot topics for their editorial content. Actually, the goal is "all about getting more hits from Google"[5]. For instance, a popular political site could benefit by looking at the current hot queries and, consequently, writing down a post on the site containing focused keywords so that Google can quickly index the post. These are just some of the many strategies adopted by SEO practitioners (Yun *et al.*, 2015).

Fortunately, the forecasting power of search query data, or perhaps we should say the computational mechanisms[6] that exploit search query data to find patterns, allow capturing quite a bit more than simple fun regularities as those just mentioned. In the following, we will introduce some of them, that, in some ways, monopolized the attention of the scientific community, and not only. Just as in the case of the interview, released to the *Financial Times*[7], by Alan Greenspan, Former Chairman of the Federal Reserve Bank. Mr Greenspan, in the middle of the subprime crisis, argued that "the essential problem is that of our models both risk models and econometric models – as complex as they have become are still too simple to capture the full array of governing variables that drive global economic reality." But what arouses more wonder is the interview released by the former governor of one of the most important economic institutions worldwide, comes immediately after, when he says "the most credible explanation of why risk management based on state-of-the-art statistical models can perform so poorly is that the underlying data used to estimate a model's structure are drawn generally from both periods of euphoria and periods of fear, that is, from regimes with importantly different dynamics." Mr Greenspan identified the weaknesses of the available models: "the underlying data used" do not seem to be able to represent "both periods of euphoria and periods of fear": probably, the underlying data desired by the banker, able to represent "euphoria" and "fear," they would have been search query data that could have better represent human behavioral dynamics since they try to capture just attention, interest, and so forth.

However, this coin, promising glory and revenues, has a dark side that arouses, rightly, the focus of the scientific community and that can be expressed with a simple question: can search queries represent public opinion of the whole population? (Zhu *et al.*, 2012). At first glance it seems that the high penetration rate of the internet (D'Avanzo and Pilato, 2015) is not sufficient to ensure that identified trends could be representative of the entire population. As supported by Zhu *et al.* (2012), "trends revealed in search queries reflect the concerns of the younger, more educated, and more active segments of the population, which are likely to be different from those of non-users." Other objections, that would require to inquiry the validity of search queries as measures of public opinion, emerge from the fact that search queries seem to measure some particular aspects of the public opinion (i.e. attention, interest, or concerns), where query trends are displayed in aggregated format, so as to preserve user's privacy. Without considering, then, the fact that active users issue more queries than less active ones, with the consequence of blinding the weight carried by each user in creating the aggregated queries.

These types of considerations have led many researchers and practitioners to question the validity of search queries as potential indicators of public opinion (Zhu *et al.*, 2012). Some others authors, more optimistically, have even gone a step further, arguing for predictive models based on search query data (Varian, 2014; Eichstaedt *et al.*, 2015) and social media sentiments data (Ruths and Pfeffer, 2014; Asur and Huberman, 2010). In our moderately optimistic, but no less enthusiastic, view, so grounded models seem, for sure, to chart the way toward a near future (Scott and Varian, 2014; Levenberg *et al.*, 2014; Leginus *et al.*, 2015; Xu *et al.*, 2014).

The growing interest in models that exploit search query data (Goel *et al.*, 2010; Moon *et al.*, 2014) and social data (Song *et al.*, 2014; Cao *et al.*, 2015; Loeb *et al.*, 2014; Karagiorgou *et al.*, 2014; Tripathi *et al.*, 2015), and their closeness with the model proposed by us in the following of the work, requires a review of the literature, although not exhaustive, about the subject matter, as it is proposed in Sections 3 and 4.

2. Forecasting real-world phenomena through search query data

Just the importance of the housing market, one of the most catastrophic aspects of the current economic crisis[8], and precisely the need to capture the dynamics mentioned by Greenspan, and cited before, that, probably, prompted Kent Webb to carry out an experiment to see if housing market search data could be used to identify trends in US home "foreclosures" (Webb, 2009). The problem is those that require quick and efficient solutions. Just think that from September 2008 to September 2012, there were approximately four million completed foreclosures in the USA; with approximately 1.4 million only in the year 2012. In this sense, the very strong correlation found between searches on the keyword "foreclosure" and actual US home foreclosures works as a source of business intelligence for organizations concerned with that market. The authors show how the typing of the query "foreclosures" on Google Trends results in the increasing of searches for this keyword at the beginning of 2008, a clear signal of the interest about the topic. Then, if we consider that Google Trends provides daily updates of weekly data, they would result, in fact, the latest data available for business intelligence purposes. The research hypothesis behind the work of Webb was:

H1. Internet searches on the keyword "foreclosure" correlate with actual US home foreclosures?

An affirmative answer to that question, together with the availability of these data, long before any other data or source, would allow search query data for "foreclosure" to be exploited by a computational early warning system for business intelligence purposes. Webb used data for total monthly US home foreclosures coming from RealtyTrac, a market research firm that releases monthly reports by state and the total US market compiled from government statistics. These data were compared with Google Trends search query for foreclosure, that, unlike RealtyTrac data, are weekly released. After an averaging of these data on monthly basis, to standardize the variables involved, the author applied an ordinary least squares regression analysis using the actual foreclosures data, released each month from January 2005 to December 2009 as the dependent variable, and the search index for foreclosures as independent variable.

The statistical analysis performed support the research hypothesis: the correlation between actual foreclosures and searches on foreclosure is strong. Simulations show that the null hypothesis of no correlation is rejected; the p-value is so small that it is presented in scientific notation. But, even more interesting, it is to observe the shape of the regression model: the search index begins to accelerate into the Fall of 2005, with the first outbreak of actual foreclosures. Search query data still overlap with the first slump in actual foreclosure happening in Fall of 2007. With the advent of 2008, search query data and actual foreclosure data show the same shape, both of them overlapping and growing over time.

American citizens, probably, right in the property bubble burst, in "fear," the same mentioned by Mr Greenspan, have directed all their attention to look for information, legislation and, in general, remedies about foreclosure, using Google or other search engines. At that time, if it had been available a model such as that offered by Webb (2009), able to demonstrate a so closed correlation among what actually happen in the world, probably, if he could not give a short-term prediction, certainly he could steer decision makers in their near future. As soon as new search query data are available, they could be added to the generated regression model so as to estimate foreclosure trends, obviously in advance with respect to the actual foreclosure data. Actually, one of the seminal works demonstrating the correlation of search query data and real-world phenomena was proposed by Jeremy Ginsberg *et al.* (2009) of the Center for Disease Control and Prevention at Google that inquired one of the major public health concerns: the epidemics of seasonal influenza. As Ginsberg *et al.* report on their own work, published on *Nature* in 2009, the World Health Organization estimates that epidemics of seasonal influenza, besides being cause of millions of respiratory illnesses, they are also responsible for many cases of death: from 250,000 to 500,000 victims worldwide each year. As a consequence, one of the major challenges in this sense would be an early detection system capable of monitoring health-seeking behavior expressed in the form of search query data on the web. The large population of web search users, as guaranteed by the penetration rate of web users (D'Avanzo and Pilato, 2015), should do the rest, in the sense that the frequency of the queries is highly correlated with the percentage of physician visits in which a patient presents with influenza-like symptoms. Reflecting the wishes of the authors, such a mechanism can achieve an accurate estimate of the current level of weekly influenza activity, with a one-day delay. In other words, such a methodology, although it is still far from being regarded as a predictive mechanism, it can surely guide public health decision makers toward the near future.

But let us look at some details of the mechanism offered by Ginsberg *et al.* As a first step, the mechanism computes time series of weekly counts for 50 million of the most common search queries in the USA. Separate weekly counts were estimated for every query in each state. Then normalization was introduced for each time series, by dividing the count for each query in a particular week by the total number of online search queries submitted in that location during the week. The proposed model would estimate the probability that a random physician visit, in a particular region, is related to an influenza-like illness. This is equivalent to the percentage of influenza-like illness-related physician visits. The probability that a random search query, submitted from the same region, is influenza-like illness related, has been used as explanatory variable by the mechanism as described in the following. The framework selects influenza-like illness search queries, requiring no prior knowledge about influenza; some of them are: influenza complication, general influenza symptoms, symptoms of an influenza complication, antibiotic medication, general influenza remedies, and so forth. A casual glance at these queries shows how they represent attention, interest or concern about the illness itself, the same that would require the intervention of a physician. This just to say that even from an informal point of view the search of such a kind of correlation model is plausible. In the following step, the authors estimated how the proposed model would fit historical data on influenza-like illness provided by the US Centers for Disease Control and Prevention (i.e. actual data). For each of the US regions analyzed they employed only a single query as the explanatory variable, rewarding queries that exhibited regional variations similar to the regional variations in influenza-like illness actual data.

Employing influenza-like illness-related query fraction as the explanatory variable, the authors fit a final linear model to weekly influenza-like illness percentages between 2003 and 2007 for all nine regions together, thus learning a single, region-independent coefficient. The model was able to obtain a good fit with actual data percentages, with a mean correlation of 0.90. The final model was validated on 42 points per region of previously untested data from 2007 to 2008, which were excluded from all prior steps. Estimates obtained a mean correlation of 0.97 with the actual data observed percentages.

Overall, Google search query data can be used to accurately estimate influenza-like illness percentages in each of the nine public health regions of the USA for which actual data by traditional surveillance sources were available. The speed with which such data can be modeled, and their resulting early availability (one or two weeks before), as well as the accuracy of such models, with respect actual data available from traditional surveillance reports, make these methodologies very interesting as early warning mechanisms also for epidemiological purposes. For instance, an early warning of search query data alerting about sharp increase of physician visits could induce public health decision makers or focus additional resources on that region to identify the etiology of the outbreak, providing extra vaccine capacity or raising local media awareness as necessary.

The case studies introduced until now show how successful can be computational mechanisms that model search query data, which are correlated with some real-world phenomena. Let us say that what is offered yet, it is not a predictive model but an exit strategy that allows us to turn toward the near future, so get out the bays of our daily lives. However, we will not have the whole story, unless we introduce those cases where more than others seem to represent a challenge to the exploitation of search query data to head toward the near future.

Even though the issues about search queries, on which we mentioned above, query data seem offering interesting perspectives with respect to public opinion polls. As Zhu *et al.* (2012) have, rightly, noted the views expressed in polls are solicited and this, as such, raises the question about validity of survey results: search users are volunteers; while survey pollsters, under the pressure of survey staff, select respondents. In contrast, search queries represent what the users, coming from some different segments of the population, are actually solicitous at the moment. The discretionary nature of search behaviors loans verisimilitude to the query data, which are not altered by search engines. Zhu *et al.* (2012), for instance, have questioned just on this chief point: to which extent the search query data represent the entire public opinion? They answered this question assessing the validity of search query data as indicator of public opinion, looking for the correspondence between them and some public interests, using public issues (e.g., housing conditions, traffic conditions, and property crimes) about the city of Shenzhen, China, thanks to the availability of actual data tracked for five years, in which opinions from both users and non-users of the internet are available for the assessment. According to the authors, Shenzhen represents the typical city in the world compared to many aspects; the internet penetration rate is about 50 percent, the population over ten million, and GPD per capita above US$15,000. Based on these aspects, reflecting the wishes of the authors, the findings of their inquiry could be easily generalizable beyond the locality, to other cities with similar characteristics.

Actual data consist of a monthly survey launched from October 2006 to December 2011, aiming at capturing Shenzhen's citizens' concerns or satisfactions about some aspects of their daily life in the city. Each survey, administered to respondents, and carried out by telephone, contains a series of questions, of which only three were being compared with the search query data, namely, those matching the search queries housing conditions, traffic conditions, and property crimes (all of them were in Chinese). With respect to the works reviewed above, that were aiming to predict real life indicators, Zhu *et al.* (2012) aim at a different purpose, namely, estimating the similarity among search query data and opinion polls. Indeed, although both variables (i.e. search query and polls) are expressed, and represented, by time series, the authors explicitly wish not remove autocorrelation from either series to hold the authenticity of the data (i.e. their endogeneity). This will allow the finding of repeating patterns, such as the presence of a periodic signal obscured by noise, or identifying the frequency in a signal implied by its frequencies. Instead, to reduce autocorrelation, the authors introduce the baseline query, using the most frequently used word in Chinese (*de* in Chinese, corresponding to "of" in English). Such a query works as "control" for some trends in query subject of matter that, for instance, could have been the product of an increasing of internet users.

The experiments consisted of repeating the regression analysis three times for each relationship among queries and responses. The first time the regression was done for the whole sample; the second time for the sub-sample of internet users (52 percent of the total sample); the third time was carried out for the sub-sample of non-users (48 percent). The overall number of regressions carried out was 30. The first pattern emerged from simulations is that some of the query series are significantly correlated with the corresponding opinion series: the resulting relationship is the strongest for the issue of property crimes and the weakest for the issue of traffic conditions. The second pattern, emerged from these simulations, shows how there is no difference among internet users and non-users. It is worth now look more closely at the three surveys issues, considering both the internet users and non-users.

As of traffic conditions, the simulations report on two search queries (i.e., jam and transportation) and a combination of them. As already said, internet users and non-users follow a similar shape of rise and fall of concerns. These shapes, however, do not match with the trends of the two queries considered, which they move, moreover, in opposite directions from each other. The question changes when the two queries are combined into one series, showing a more parallel shape with survey's responses.

As of the housing conditions issue, the "housing" search query does not follow the corresponding public opinion expressed through the survey, which responses appear more fluctuating. In this case, internet users and non-users of the survey appear to behave very different each other. An interesting aspect is represented by the parallelism between the search query and the baseline query, meaning that the increasing of the searching for the query "housing" is chiefly due to the increase of the number of searches happening in the same period. Looking at the search query "housing price," all three trends (i.e. search query, user, and non-user opinion poll) follow the same shape, with two cycles of ups and downs that show a non-casual pattern.

As said before, another public issue matter of investigation was property crimes. The corresponding opinion polls were compared with three keywords: thief, stealing, and robbery. Among the three keywords only "stealing" is more resembling to the curve representing internet users' and non-users' opinion polls. The keyword thief behaves the same way as housing, moving parallel to the baseline query, showing, as consequence, how it probably is due to the increasing of internet searches, as a result of some events not related (e.g. Google's removal of search engine servers from Mainland China[9] to Hong Kong).

Overall, opinion polling is expensive, both humanly and in economic terms, and shows a rising refusal rate. From them, search queries do capture concerns of the general public, but showing a fluctuating degree of accuracy across different issues. Even if representative of the general population, search queries seem to be able of estimating only specific aspects of public opinion (e.g. attention), leaving out other relevant aspects cognitive and social aspects of public opinion that seem to be better captured by other social media such as social networks, microblogs, and the like.

What seems to emerge is the need for an experimental framework that is capable of giving the experimenters the ability to test models and hypotheses in a systematic way, integrating both search queries data and social data. While on the exploitation of the former we just reported above, in next Section it is introduced a short and recent review where the latter is employed.

3. Forecasting real-world phenomena through social media sentiments

Compared to traditional blogs, microblogging has shown a tremendous growth in popularity in recent years. Microblogging users upload new messages more frequently and instantly. One of the most representative examples is Twitter because of its popularity and data volume. Currently, more than 500 million users around the world are using it to share information, opinions, news, moods, concerns, facts, rumors, and general events of public interest as earthquakes, political events, deaths of famous people (Liu *et al.*, 2014; Ceron *et al.*, 2014; Middleton *et al.*, 2014; Burnap *et al.*, 2015; Ahn and Spangler, 2014). Corporations use Twitter to make announcements of products, services, and events (Schivinski and Dariusz, 2014). Twitter is also a news media: many news outlets in fact have accounts on Twitter to report news. Tweets can be seen as a source of data enabling users and corporations to stay informed of what is happening now or what is being said about them (Terrana and Pilato, 2013).

Twitter is increasingly exploited for other various research tasks, including modeling and predicting users' behavior (Goel *et al.*, 2010; Ruths and Pfeffer, 2014).

However, one of the fields of research that has most attracted the scientific community was predicting election results (Tsakalidis *et al.*, 2015; Scott and Varian, 2014; Burnap *et al.*, 2015; Gayo-Avello, 2013). A comprehensive review is not the purpose of this discussion. However, as in the next section we will show how to use Twitter social data coming from the political debate, in this context we will own Twitter sentiments to predict election outcomes. Gayo-Avello (2013) offers an in interesting point of view of the subject matter, supporting that the "predictive power regarding electoral prediction has been somewhat exaggerated." According to the author, "social media may provide a glimpse on electoral outcomes but research has not provided strong evidence to support it can currently replace traditional polls." As we have said at the end of the previous section, search query data, on their own, did not have the power to replace actual data (e.g. poll data). And in turn, not even social data have the strength to replace the poll data and/or search query data. In our view, all these data, as a whole, should be integrated in order to provide more and better models of social phenomena, in the round. This view is known as evidence integration (Williamson, 2008). In other words, the final model must integrate the totality of the available evidence coming from different data sets (e.g. search query data, social data, polls, etc.).

We will focus on the work done by Tsakalidis *et al.* (2015), which released the results of its analysis a few days before the results, showing a high degree of precision with regard to Greece's election results.

The paper Tsakalidis *et al.* (2015) offers one of the most recent and complete attempts of exploiting Twitter's content and polls in order to predict EU election results in different countries. Treating users' voting intentions as time-variant features, the authors look at Twitter political discussions in order to create an overall index of which vary with the time, without predicting each user's vote. Then, for every party, 11 Twitter features are identified and combined with one poll-based feature using a multivariate time-series model. For instance, the number of tweets mentioning a certain party on a specific day is a Twitter-based feature, while the percentage of that party reported on a poll that was conducted on that day is a poll-based one. As a next step, these features are fed to different forecasting algorithms so to predict every party's voting share, independently of each other. Worth, at this point, take a look at some details of the approach just introduced. First, since many Twitter-based features are sentiment related, each of them was labeled with its own sentiment, adopting a lexicon-based approach: SentiWordNet[10], containing 150,000 synsets with a double value indicating their polarity; Opinion Lexicon[11], containing 6,800 polarized terms; and the Subjectivity Lexicon[12], containing 8,000 terms along with their POS, subjectivity – strong/weak, and polarity indication. They were assigned the values of 1 and ?1 for the positive and negative terms in the Opinion Lexicon; while for the Subjectivity Lexicon, they used four values (−1, −0.5, 0.5, 1) to represent every subjective word based on subjectivity (|0.5| for weak, |1| for strong) and polarity; finally, for SentiWordNet, they kept the values of every synset.

As for Twitter data, they have aggregated 361,713 tweets from 74,776 users in Germany; 452,348 from 74,469 users in the Netherlands; and 263,465 from 19,789 users in Greece. A general pattern emerged showing that negative opinions dominate in political discussions (−0.54 for Germany, −1.09 for the Netherlands, and −0.29 for Greece). Instead, as regards opinion polls data they were employed 26 different polls from 11 diverse sources in Greece; nine polls from four sources in Germany; and 13 polls from three sources in the Netherlands[13].

All together, combinations of these specific party's features (11 Twitter- and 1 poll-based) have been fed as input to three different algorithms: linear regression, Gaussian process, and sequential minimal optimization for regression.

Evaluations performed, using the standard mean absolute error, show fluctuating results. Authors compared their methods with respect to past works, commercial solutions, and combination of them. For instance, according to the authors, in Germany polls are the second best predictor. While both in Greece and the Netherlands they had a lower performance with respect to combination of poll-based methods. From authors' point of view, when polls data are used as training for their own proposed algorithms (i.e. linear regression, Gaussian process, and sequential minimal optimization for regression), the latter outperform polls. However, although the proposed models perform better in terms of error rate terms, they do not work well in terms of correct ranking of the parties.

We are still in the presence of some fluctuations with respect to the expectation about this model. As well as search query data have a complementary role, compared to the poll data, in the same way social data have, in turn, a complementary role than the two previous. The role of the experimental framework introduced in next Section is to have a platform that allows us to integrate the evidence of a different nature in our possession (i.e. search query data, social data, poll data, etc.) to test hypotheses able to model the phenomena under investigation.

4. Social nowcasting: an experimental framework

The framework, whose architecture is depicted in Figure 1, makes use of two kinds of analysis: the first one is aimed at measuring the sentiment, while the second one is oriented at estimating the emotions expressed in posts broadcasted on social networks. In order to find the hot topics and the most relevant related queries to be issued on Twitter, we exploit Google Trends, which summarizes queries through the analysis of web users search behaviors. Roughly speaking, Google Trends it is a tool designed for tracking the popularity of any given search term over time (Dzielinski, 2012). Google Trends provides an index of the volume of Google queries by geographic location and category. Its data do not report the raw level of queries for a given search term. Rather, it reports a query index. The query index starts with the query share: the total query volume for search term in a given geographic region divided by the total number of queries in that region at a point in time. Query share numbers are then normalized so that they start at 0 in January 1, 2004. Numbers at later dates indicated the percentage deviation from the query share on January 1, 2004. The user selects the category and the world region of interest; after that she selects the story to be explored, and Google Trends shows her a page reporting a title characterizing the clicked story, the most relevant articles related, a chart showing both the

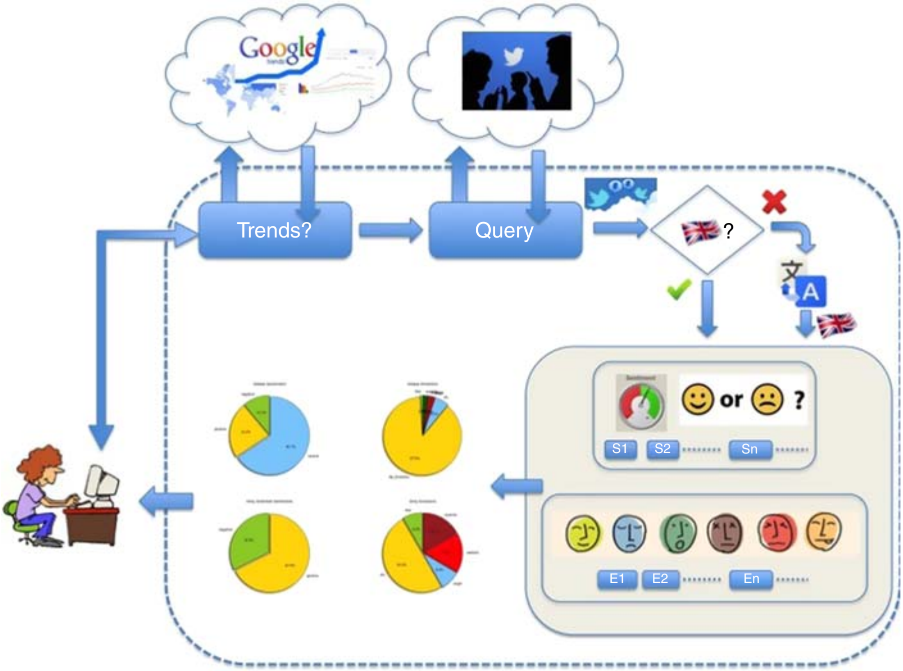


Figure 1.
An overview of system

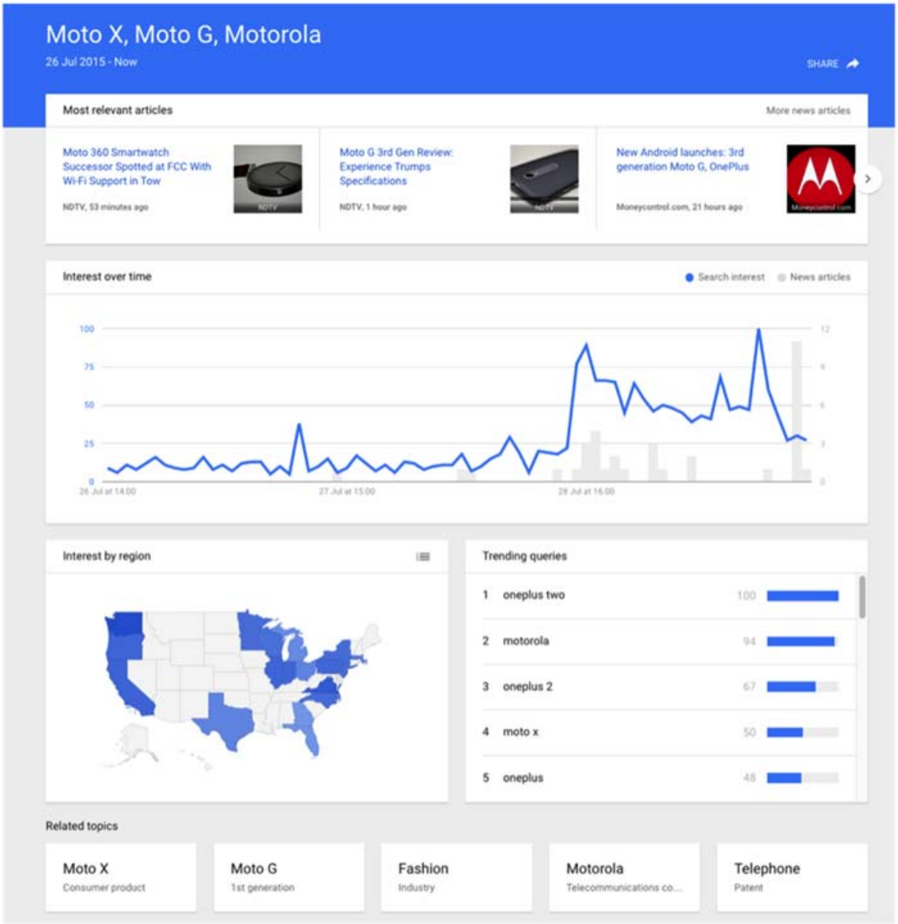
search interest and the news articles published on the web, the interest by region, and the trending queries. At the end of its page, Google Trends shows also the related topics, as shown in Figure 2.

Subsequently, the user selects the suggested queries and the system exploits them for questioning Twitter for the most recent published satisfying the queries.

A user's action allows the initialization of the framework that starts with a call to Google Trends, in order to explore, and/or select, the trending stories belonging to a category of interest, or to the top stories. Once the story of interests is chosen she looks for the corresponding trending queries, which are then submitted to the following module; the user, at this point, can determine the maximum number of tweets desired for each query.

Tweets are retrieved by using Twitter APIs[14], exploiting the default access level, known as spritzer, returning, randomly, a small proportion of all public tweets (i.e. 1 percent). By using a special account, Twitter APIs can also provide two other levels of access, the firehose and the gardenhose, returning, respectively, 100 and 10 percent of all public tweets.

The retrieved data, returned as a set of JSON objects[15], in addition to the text of the tweet, contain other features, such as date, source, type, profile, location, number of favorites friends, followers, URL, hashtag, and so forth. For the experiments presented below were considered only the timestamp and the text for each tweet. All other types of information, just mentioned, are used for other experiments in course of work. The language of each tweet is automatically detected and, if it is different from English, an online translation service is invoked, so as to translate in the best possible way the currently examined tweet using only English. In fact, according to Mohammad *et al.* (2016), Do and Choi (2015) and Sangiorgi *et al.* (2013), in some cases a translation procedure can be useful for detecting sentiment in language other than English.



Moto X, Moto G, Motorola

26 Jul 2015 - Now

SHARE

Most relevant articles

More news articles

Moto 360 Smartwatch Successor Spotted at FCC With Wi-Fi Support in Tow

NDTV, 53 minutes ago

Moto G 3rd Gen Review: Experience Trumps Specifications

NDTV, 1 hour ago

New Android launches: 3rd generation Moto G, OnePlus

Moneycontrol.com, 21 hours ago

Interest over time

● Search interest ● News articles

Interest by region

Trending queries

1	oneplus two	100
2	motorola	94
3	oneplus 2	67
4	moto x	50
5	oneplus	48

Related topics

Moto X Consumer product	Moto G 1st generation	Fashion Industry	Motorola Telecommunications co...	Telephone Patent

Figure 2.
The Google Trend page for "Moto X, Moto G, Motorola"

At this point of the development of the entire framework, this allows us to refine as much as possible the modules as a function of only one language.

All tweets so obtained are, then, preprocessed as in the following: stop words are filtered out, links and hashtag are removed, and words of length less than three characters were discarded before processing the text because they often hide off-topic posts or even spam. The tweets containing mainly abnormal sequences of characters were discarded (Terrana and Pilato, 2013).

The processing step just described, culminates in the intervention of the two modules of sentiment and emotion analysis. In particular, the sentiment detection module estimates the predominant orientation, i.e. positive or negative, of each tweet. Simultaneously, the emotion detection module identifies the emotions expressed by the current tweet, giving as a result the predominant feeling among one of the following: anger, disgust, fear, joy, sadness, and surprise (Strapparava *et al.*, 2006).

The choice of these six emotions comes from the psychological evidence of human non-verbally expressed emotions proposed by Ekman (1992).

All the modules of the architecture are tunable at will for two orders of reasons. First, as the entire framework is used to formulate hypotheses (Spangler *et al.*, 2014; Kitchin, 2014;

Tang *et al.*, 2014), in different fields of application, constantly, the different modules are enriched with new features (e.g. dictionaries, natural language processing routines, learners, and so forth), and, as such, they must be able to be replaced easily. Second, the general wish is to have in hands an experimental framework that can be as reliable as possible; therefore we believe that this architectural choice, it will take us in the right direction.

At this point, we introduce some details about both the sentiment and emotion detection modules that, at this stage of overall development of the framework, cover the most important experimental role. Furthermore, some aspects of the visualization module are also introduced.

4.1 Sentiment detection module

Sentiment detection module exploits different sentiment detection tools, constituting a sub-module, which can be plugged or unplugged, at will. In particular, at present the sentiment detection module exploits four sentiment analysis tools: a naïve Bayes detection algorithm, a simple voter algorithm, an NLTK-based sentiment analysis algorithm[16], and a commercial sentiment analysis tool[17]. All of them are briefly described as follows:

1. The data set contains 6,518 terms, each of which is associated to a sentiment which can be "positive" naïve Bayes detection algorithm has been trained on Wiebe's subjectivity lexicon (Wilson *et al.*, 2005); overall, or negative." Furthermore, a weight determines if the term is strongly or weakly subjective. In total, 2,324 are positive terms, of which 1,482 strongly subjective and 842 weakly subjective; 4,176 are negative terms, with 1,097 weakly subjective, 3,079 strongly subjective; and 18 terms can be associated either to positive or negative subjectivity, of which 16 are strongly subjective and two are weakly subjective. The learning module analyzes a given text and for each polarity returns its absolute log likelihood expressing that sentiment; results are then evaluated, resulting in the most likely polarity.
2. The voter algorithm uses the above-mentioned lexicon, simply counting the number of occurrences of the positive and negative words contained in the text.
3. An NLTK-based tool for sentiment analysis that was preventively trained on a movie reviews corpus has been included.
4. Finally, a commercial sentiment analysis tool has been also included in the sentiment detection module.

The outcome of each sub-module is the percentage of retrieved tweets classified as expressing a positive, negative or neutral. Finally, the results from all the submodules on the retrieved tweets are averaged, obtaining a distribution of positive *Pos* 2 [0,1] negative *Neg* 2 [0,1], and neutral *Neu* 2 [0,1] orientation, where of course is $Pos+Neg+Neu=1$.

4.2 Emotion detection module

As well known, Tweets can express also emotions and, as such, this module estimates the most appropriate one for each tweet among the six basic emotions: anger, disgust, fear, joy, sadness, surprise proposed by Ekman (1992). The emotion detection module can exploit different emotion detection tools, embedded as submodules, which can be plugged or unplugged at will. In particular two submodules, a nave Bayes learning algorithm and a voter algorithm have been at present plugged in the emotion detection module. Each of them is briefly summarized as follows:

- The learning algorithm is trained on the emotions lexicon provided by Strapparava *et al.* (2006); this data set contains 1,542 terms, each of which is associated to one of the six above-mentioned emotions: 355 terms for anger, 70 for disgust, 195 for fear, 553 for joy, 274 for sadness, and 95 for surprise.

- The simple voter algorithm uses the above-mentioned lexicon by counting the number of occurrences of the anger, disgust, fear, joy, sadness and surprise words contained in the text. The majority of counts gives the prevalent emotion associated to the text message. If the text does not contain a prevalent number of words expressing a given emotion, the text message is labeled as carrying "no emotion."

The outcome of each sub-module is the percentage of retrieved tweets classified as expressing an emotion among the aforementioned six or classified as "no emotion." Finally, the results coming from all the submodules on the retrieved tweets are averaged, obtaining an "emotion" distribution, obtaining:[0,1], disgust 2 [0,1], fear 2 [0,1], joy 2 [0,1], sadness 2 [0,1], surprise 2 [0,1] and n*o* emotion anger 22 [0,1], whereas the overall sum must be equal to 1. As is the case for the sentiment module, results obtained from all the retrieved tweets are averaged, and the averaged distribution is the outcome of the module.

4.3 Visualization module

An overall result is therefore presented to the user in a graphical manner, in order to help her in her decisional process (Dhar and Simonson, 2003). Data visualization offers to decision makers a way to make sense of large data set, allowing the discovering of patterns for decision support (White, 2011).

The user can decide also to plot and compare the analysis results regarding different queries in order to get an idea on the general feelings that arise from twitter social network regarding specific themes or features of interest. This feature can be helpful also for the comparison of sentiments and emotions arising from tweets retrieved by using different queries.

This approach simplifies the decisional process and allows overcoming the information overload by quickly having an idea about the general sentiment or emotions raised by news goods or aspects. In the following section, we give a couple of examples regarding the system use.

5. Experiments: nowcasting trending issues

We have implemented and tested the prototype employing different queries arising from Google Trend. In the following, we report two examples regarding the keywords. The examples that we report mainly refer to the consumer market. The choice is dictated by the exhibition which aims this context, and by similar experiments, using similar products, we proposed elsewhere (Dhar and Simonson, 2003). In fact, as we have said, the framework proposed represents for us an experimental laboratory where we can test hypotheses and models on different social phenomena, using social behavioral data. Examples reported are as follows:

- Moto X, Moto G, Motorola.
- New York Legionnaire.
- Republican debate.

Each of them is dedicated to one of the following sections.

5.1 Case study 1: Moto X, Moto G, Motorola Car

The example regards the Trend Story entitled "Moto X, Moto G, Motorola," shown in Figure 2. In this case the most relevant articles associated to this story, as reported in Figure 2 are as follows:

- Moto 360 Smartwatch Successor Spotted at FCC with Wi-Fi Support in Tow ("Motorola on Tuesday launched the Moto G (Gen 3) at a launch event in India, while it revealed the Moto X Play and Moto X Style smartphones at a separate event. Many had expected that the company would announce the second-generation Moto 360 on the same day; though that didn't happen [...]").

- Moto G 3rd Gen Review: experience Trumps Specifications ("The smartphone industry has a specifications problem and the tech press must share a large part of the blame. We are living in a world that's obsessed with octacore processors, 13-megapixel cameras, and gigabytes of RAM, but we don't spend enough time talking about the thing that really matters – how it feels to use a phone on an everyday basis").
- New Android launches: 3rd generation Moto G, OnePlus ("The smartphone war has two new con-tenders. Motorola and OnePlus unveiled their new offerings in New Delhi Tuesday. Here are the first impressions of the new Moto G and the OnePlus 2!").

According to what reported in Google Trend, the Query module asked Twitter for at most 500 tweets on the following queries: "moto x," "moto g," "motorola," "oneplus two," "oneplus 2," "oneplus." We retrieved 500 tweets for each query and the time interval was from "Wed Jul 29 14: 13: 56 +0000 2015" to "Wed Jul 29 14:18:08 +0000 2015." The results are graphically summarized in Figures 3 and 4. Figure 5 reports the percentage of tweets that have a sentiment orientation or not, and the percentage of the same tweets that express an emotion or not. As it can be seen, the queries "Motorola" and "moto x" present the highest percentage of oriented tweets (42.9 and 39.4 percent, respectively), while the queries "moto g," "moto x," "Motorola," and "oneplus" present the highest percentages of tweets with an emotion (10.1, 10, 9.1 and 8.1 percent, respectively). Looking at Figure 6, it can be observed that for all the queries there is a strong positive orientation, an average of 86.3 percent with a peak on the 97 percent for the "oneplus two" query, and there is a very low percentage of negative sentiment: an average of 13.5 percent. For what concerns emotions expressed in the retrieved tweets, we obtain an average value of 48.3 percent of "joy" followed by an average of 24.3 percent of "surprise." This means that probably there has been a strong positive and joyful reaction to this trending story highlighted by Google Trends. This can lead to the decision of the user to buy or at least to follow the evolution of this product for a future purchase of it.

5.2 Case study 2: NYC Legionnaires' disease outbreak

Looking at Google Trends our attention was caught by the query "new york Legionnaire," related to the recent outbreak of Legionnaires' disease in New York City.

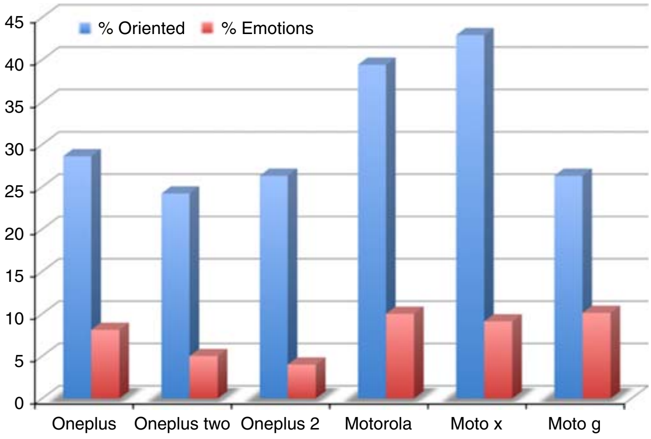


Figure 3.
Percentage of oriented tweets and tweets carrying an emotion among those retrieved for each query

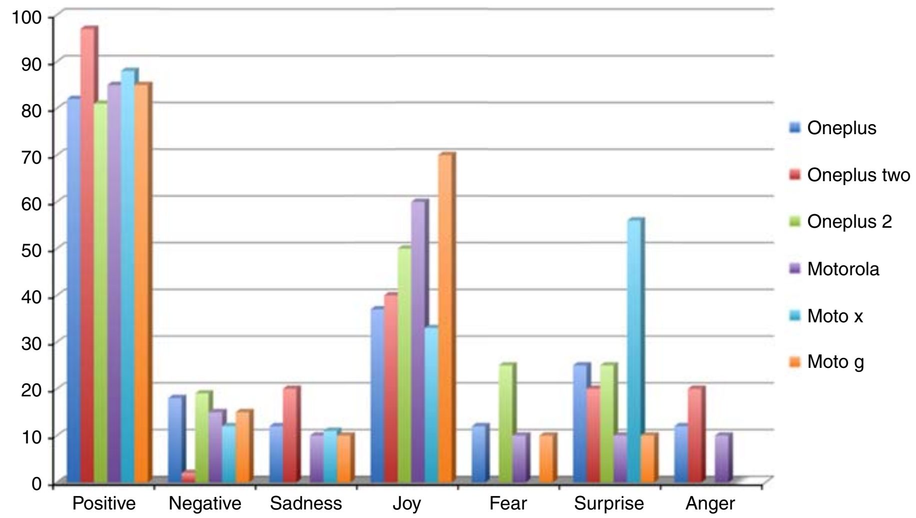


Figure 4.
Percentage for each query of positive or negative orientation and emotions among oriented tweets or tweets carrying an emotion

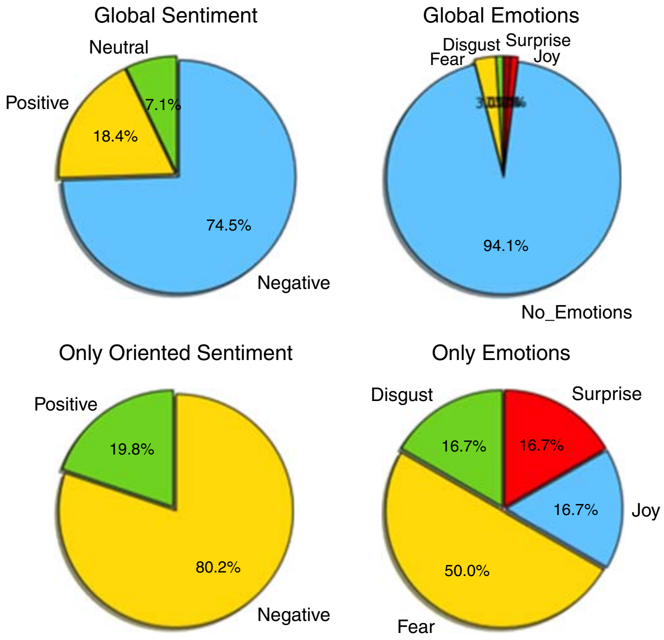
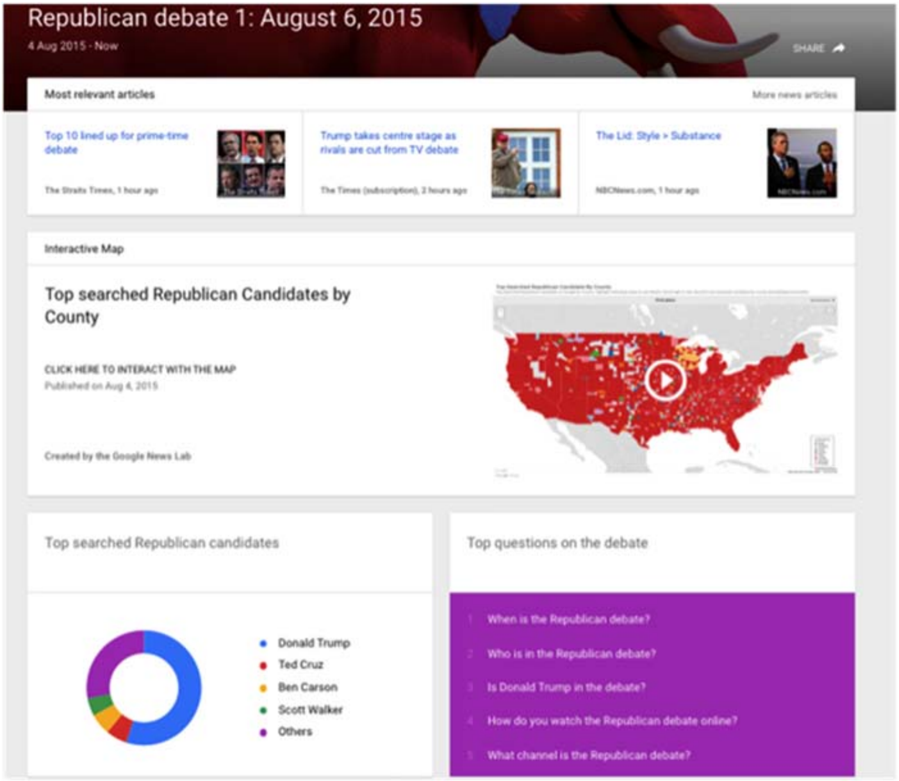


Figure 5.
Results for the "New York Legionnaire" query

As a consequence, we have set our system for querying Twitter with the aforementioned keywords. In Figure 5, we show the clear outcome of the system obtained by automatically examining 300 retrieved tweets between Wed Aug 05 20:14:09 +0000 2015 and Thu Aug 06 21:28:31 +0000 2015.

As it can be seen there is a value of 74.5 percent of retrieved tweets having negative sentiment (e.g. Thu Aug 06 21:14:41 +0000 2015 - b"Legionnaire's Disease Most Deadly for



Republican debate 1: August 6, 2015

4 Aug 2015 - Now

SHARE

Most relevant articles

More news articles

Top 10 lined up for prime-time debate

The Straits Times, 1 hour ago

Trump takes centre stage as rivals are cut from TV debate

The Times (subscription), 2 hours ago

The Lid: Style > Substance

NBCNews.com, 1 hour ago

Interactive Map

Top searched Republican Candidates by County

CLICK HERE TO INTERACT WITH THE MAP

Published on Aug 4, 2015

Created by the Google News Lab

Top searched Republican candidates

- ● Donald Trump
- ● Ted Cruz
- ● Ben Carson
- ● Scott Walker
- ● Others

Top questions on the debate

1. When is the Republican debate?
2. Who is in the Republican debate?
3. Is Donald Trump in the debate?
4. How do you watch the Republican debate online?
5. What channel is the Republican debate?

Figure 6.
Google Trends page for "Republican debate 1"

Frail, Elderly, Experts Say: THURSDAY, Aug. 6, 2015 – As New York City … http://t.co/ZayAlqmDas"), and the prevalent emotion is "fear" (e.g. Thu Aug 06 11:52:18 +0000 2015 - b'Daughter of Legionnaires victim horrified at NYC outbreak: A Manhattan woman whose mother died of Legionnaire […] http://t.co/8GuDlwvcKV') This result, for example, can help the users to postpone a possible travel to NYC or to take specific precautions for preventing the disease.

5.3 Case study 3: temporal trends in social sensing, the Republican debate

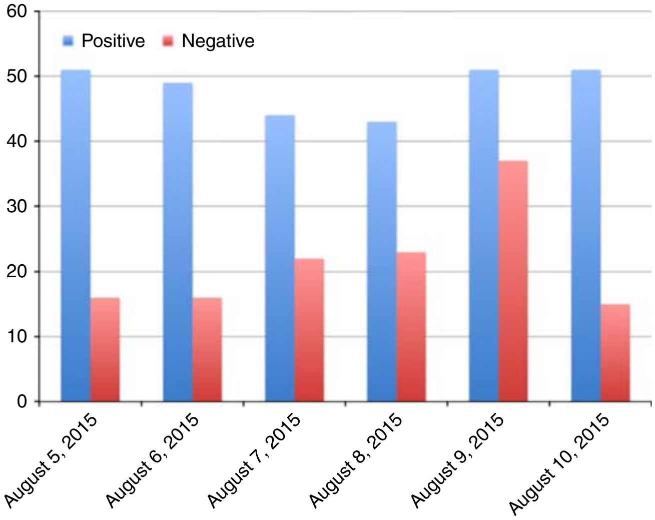
In this section, we illustrate the temporal trend obtained monitoring Twitter for five days with regard to the republican debate and the Nashville shooting.

5.3.1 Republican debate. In this subsection, we analyze the results obtained by monitoring the "republican debate" "hot topic" as reported by Google Trends (see Figure 6). In particular, we have focused our attention to the four top searched Republican Candidates reported by Google, namely, "Donald Trump," "Ted Cruz," "Ben Carson," and "Scott Walker."

We queried Twitter once a day for each of them, retrieving a maximum of 300 tweets for each candidate. At the end of the five days we have plotted the trends of positive and negative sentiments and the trends of the six fundamental emotions in order to better catch the feelings arising from the Twitter stream regarding the aforementioned candidates.

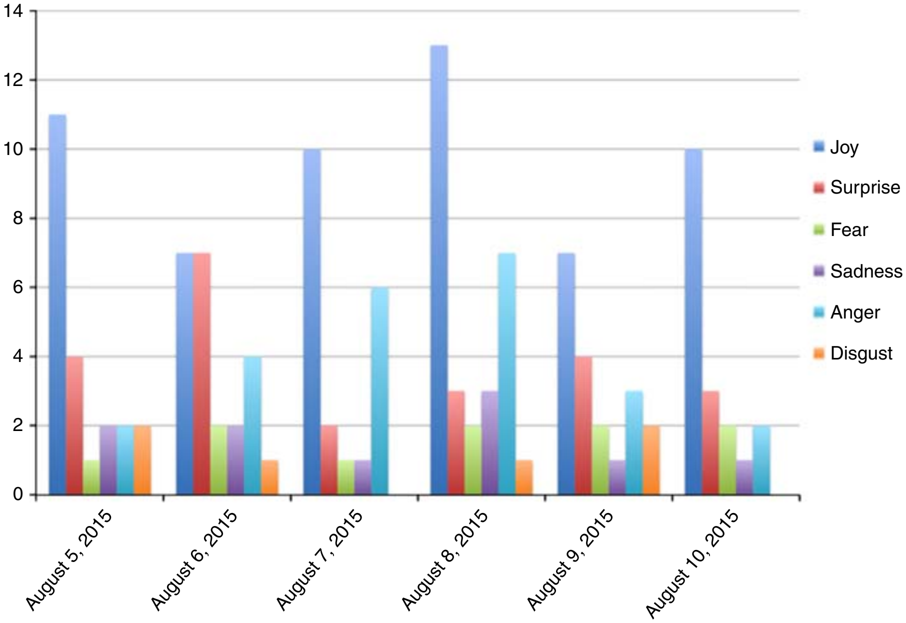
In the following we present the obtained results:

(1) Donald Trump: For the candidate "Donald Trump" the charts related to the sentiment and the emotions detected among 300 retrieved tweet are reported in Figures 7 and 8.



Note: Percentages on a sample of 300 tweets retrieved

Figure 7.
Polarity chart of tweets regarding Donald Trump



Note: Percentages on a sample of 300 tweets retrieved

Figure 8.
Emotion chart of tweets regarding Donald Trump

From Figure 7 we can observe that there is a growing percentage of negative posts and a percentage of positive posts that ranges between 43 and 52 percent. We have a peak of both positive and negative feelings on August 9. In the following we report some of the tweets that let arise these peaks:

Negative sentiment

- Sun Aug 09 16:03:27 I was talking about Megyn Kellys period is a deviant'+0000 2015 - b'http://t.co/Kaw9NThDAK Donald Trump: Anyone who thinks
- Sun Aug 09 16:03:12 about Megyn Kelly roils GOP race 307 Ontario CAN Mississauga CAN http:+0000 2015 - b"RT @smitharyy: Megyn Kelly #MegynKelly Trump's comment//t.co/VKTym"
- Sun Aug 09 16:02:57 +0000 2015 - b'RT @FoxNewsSunday: Sen. Rand Paul (R-KY), Presidential candidate, doubts Donald Trump's conservatism: "He really could be a liberal[…] http"

Positive sentiment

- Sun Aug 09 16:03:48 saved us from another depression. http:+0000 2015 - b'Sept. 2010, Trump said he did agree that President Obama had//t.co/dYq5pfB64v'

For the emotions we have a peak of the feeling "surprise" on August 6 and 9. The peak of "surprise" on August 6 is mainly due to 12 tweets like this:

- Thu Aug 06 20:54:28 Donald Trump mentions Mexico tonight http: + 0000 2015 - b'RT @Eater: How to get free tacos from @Eat24 every time//t.co/QoJcHiOZV1 http://t.co/EcoPLU7fYv'

Other message classified as "surprise" on the same day is:

- Thu Aug 06 20:53:31 +0000 2015 - b'Donald Trump is starting to get my attention'

Another predominant feeling on August 6th is "joy":

- Thu Aug 06 20:56:07 debating together its like meek and drake but better'+0000 2015 - b'RT @BENINISM: imagine donald trump and bernie sanders
- Thu Aug 06 20:55:03l and, can you tell Looks like a 757. @wkyc #GOPDebateCLE #GOPDebate http:+0000 2015 - b'RT @WKYCAndrewH: Donald Trump has landed in Cleve//t.co/7UGq1g2XWW'-

There is a peak of joy on August 8, and these are some of the relevant tweets labeled by the system as "joy":

- Sat Aug 08 15:10:20 his gaffes are legitimately making me die laughing'+0000 2015 - b'Donald Trump is the biggest joke in the history of America, like
- Sat Aug 08 15:10:13 is only one man he is Donald Trump @mcuban #TRUMP2016 http:+0000 2015 - b'What the loyal American people need is a great leader and there//t.co/r0wu5gUQx2'
- Sat Aug 08 15:10:37 Kelly remark gt; love that replacement speaker is Megyn Kelly http:+0000 2015 - b"Conservative Forum Rescinds Trump's Invitation Over Megyn//t.co/g7UWwqjzG6"
- Sat Aug 08 15:09:36 vastly expanded legal immigration, and Bernie… http:+0000 2015 - b'Love to see Bernie and Trump on one ticket. Trump options for//t.co/dxeesRCb3k'

- Sat Aug 08 15:10:49 lashing – don't mess w their @RealDonaldTrump http: + 0000 2015 - b"RT @Lrihendry: 2 women give Megan Kelly hilarious tongue//t.co/nGBLp66cJb http://t.co/"

The "surprise" feeling of August 9 is mainly due to seven tweets like:

- Sun Aug 09 16:04:11 Representative Makes A Surprising Statement About Donal http:+0000 2015 - b'#News #Retweet #Retweet Joaquin El Chapo Guzmans Mexico//t.co/dLFsm51dEd'

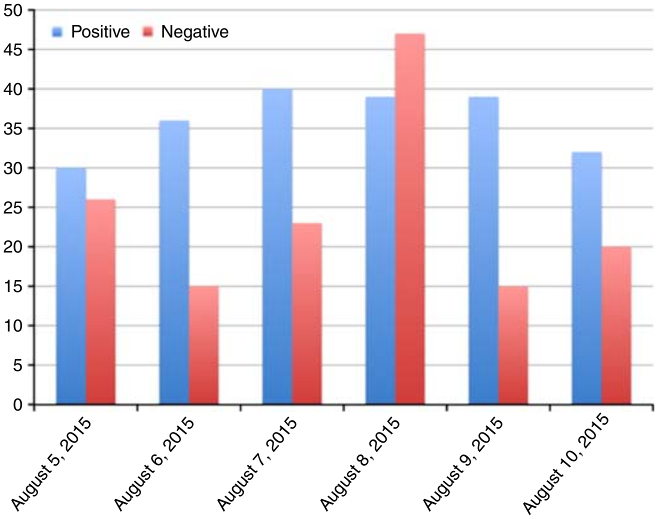
Other tweets that let arise "surprise" are:

- Sun Aug 09 16:02:38 candidate Donald Trump defends his relations with wome… http:+0000 2015 - b" 'Women are tremendous' says Trump - Republican presidential//t.co/CziLCmcl6g"

Anger is another predominant feeling on August 8th, arising from tweets like:

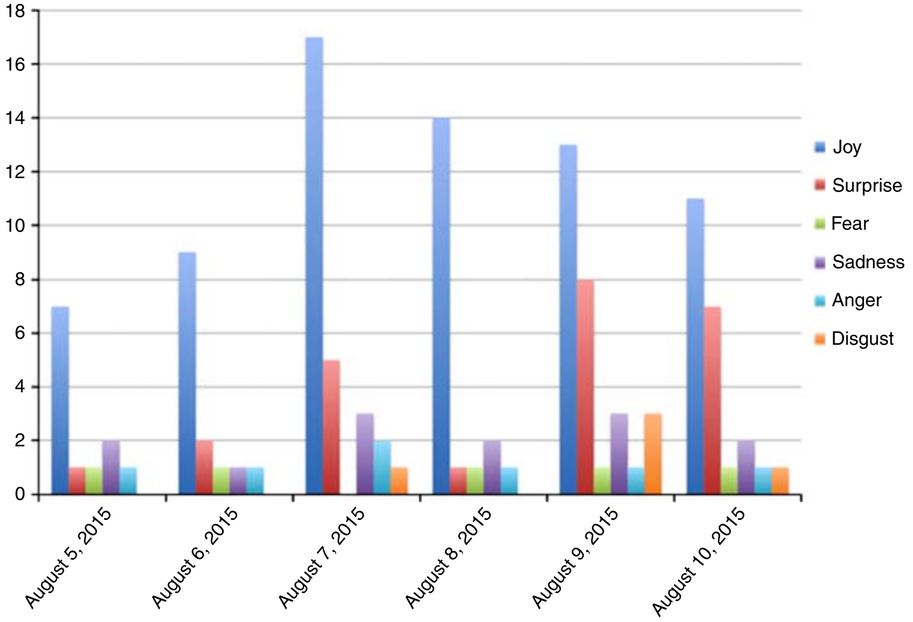
- Sat Aug 08 15:10:16 second that @EWErickson is any less the hateful sexist than is Donald Trump http:'+0000 2015 - b'RT @Max Fisher: Lets not pretend for even 1/100th of one Sat Aug 08 15:10:13 +0000 2015 - b'RT @drmoore: I totally agree with my friend @EWErickson standing up to @realDonaldTrump misogyny: http://t.co/tvK0mk1mx3'
- Sat Aug 08 15:10:03 hostile female journalist was menstruating http:+0000 2015 - b'Donald Trump barred from Republican event after implying//t.co/hXsDs9uJqm'
- Sat Aug 08 15:09:42E verything Trump said about you is correct. Thank you Donald Trump for exposing evil'+0000 2015 - b'@megynkelly your are biggest most evil anti-muslim bigot.

(1) Ted Cruz: for the candidate "Ted Cruz" the charts related to the sentiment and the emotions detected among 300 retrieved tweet are reported in Figures 9 and 10.



Note: Percentages on a sample of 300 tweets retrieved

Figure 9. Polarity chart of tweets regarding Ted Cruz



Note: Percentages on a sample of 300 tweets retrieved

Figure 10.
Emotions chart of tweets regarding Ted Cruz

For what concerns Ted Cruz we have a negative sentiment peak on August 8th, arising from tweets like the following ones:

- Sat Aug 08 15:07:28 Ted Cruz? http://t.co/z2d5jk1MG6 https:+0000 2015 - b'Why do you call @donlemon ignorant, but not white Senator//t.co/Azl4AnJ4Rl' Sat Aug 08 15:05:30 +0000 2015 - b'RT @classenchandler: "I would destroy ISIS by pulling out my red rider BB gun and shoot some Muslims? -Ted Cruz #GOPDebate'
- Sat Aug 08 14:59:04 peachobama http://t.co+/VrhjgOXTQU"0000 2015 - b"Ted Cruz Lists 76 'Lawless' Obama Actions - Breitbart #im-
- Sat Aug 08 14:48:07" radical Islamic terrorism." Beetle juice strategy! #GOPDebate'+0000 2015 - b'RT @JillFilipovic: So Ted Cruz will destroy ISIS by […] Saying

The same day is characterized also by positive sentiments, arising from 37 tweets similar o equal to:

- Sat Aug 08 15:08:52 dom, Liberty, and Dictators USA http: +0000 2015 - b'RT @brownjenjen: #TedCruz Ted Cruz Ted Cruz Loves Free//t.co/yRINjawGDs http://t.co/BkU1bgT6Na'-

For what concerns emotions there is a peak of joy on August 7, arising from 13 posts like:

- Fri Aug 07 20:14:27 vatives to our Tennessee Leadership Team: … http:+0000 2015 - b'Ted Cruz @tedcruz Very glad to add these courageous conser//t.co/yso9XZldJl'-

And also from the following tweets, automatically labeled by the system with the label "joy":

- Fri Aug 07 20:13:00 Egypts dictator." http://+t.co0000 2015 -/OEVrCcSISB http:b'RT @Slate: Ted Cruz: "Our president should be more like//t.co/5vLH2oIT3o'

- Fri Aug 07 20:05:10http://t.co/vYk50SCzmr"+0000 2015 - b" If you're a Ted Cruz fan, this will make you love him even more!

Surprise is another predominant feeling on August 7, arising from seven posts like:

- Fri Aug 07 20:14:11 rude,condescending remark she made was to Ted Cruz about God,can you get anymore disr'+0000 2015 - b'RT @GARock945: Forget the Trump/ Megyn rumble,the most

And also from posts like:

- Fri Aug 07 20:19:39 Ignorance is amazing. Ted Cruz.' +0000 2015 - b'@RayRay180 @tedcruz Birthers are out in force today. Their

The peak of surprise on August 9 is mainly due to ten tweets like:

- Sun Aug 09 16:07:09 Megyn Comments http: +//0000 2015 -t.co/bBOo6QsHoo Hope It Doesn't Get Him Banned Fr Debates http:b"RT @LeahR77: Classy Cruz Refuses 2Slam Trump Over//t."

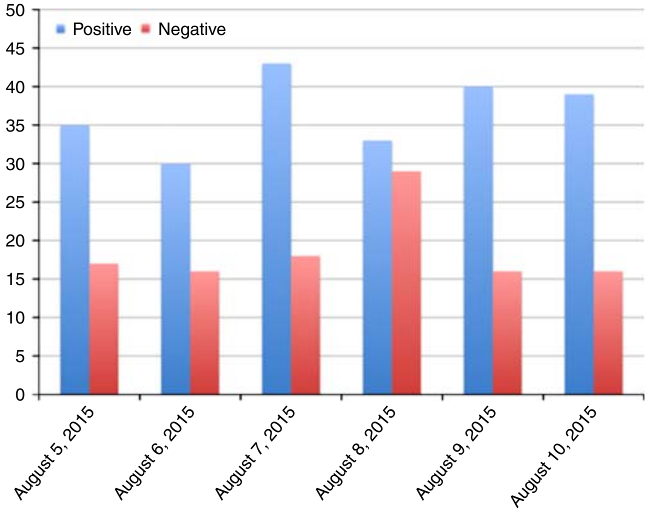
And five tweets like:

- Sun Aug 09 16:00:04 Senate Leadership Tells Ted Cruz To Get Lost http:+0000 2015 - b'RT @Champergirl: CRONY CAPITALISM AT ITS FINEST//t.co/32kUoqlbGI #WakeUpAmerica #tcot'!

(1) Ben Carson: for the candidate "Ben Carson" the charts related to the sentiment and the emotions detected among 300 retrieved tweet are reported in Figures 11 and 12.

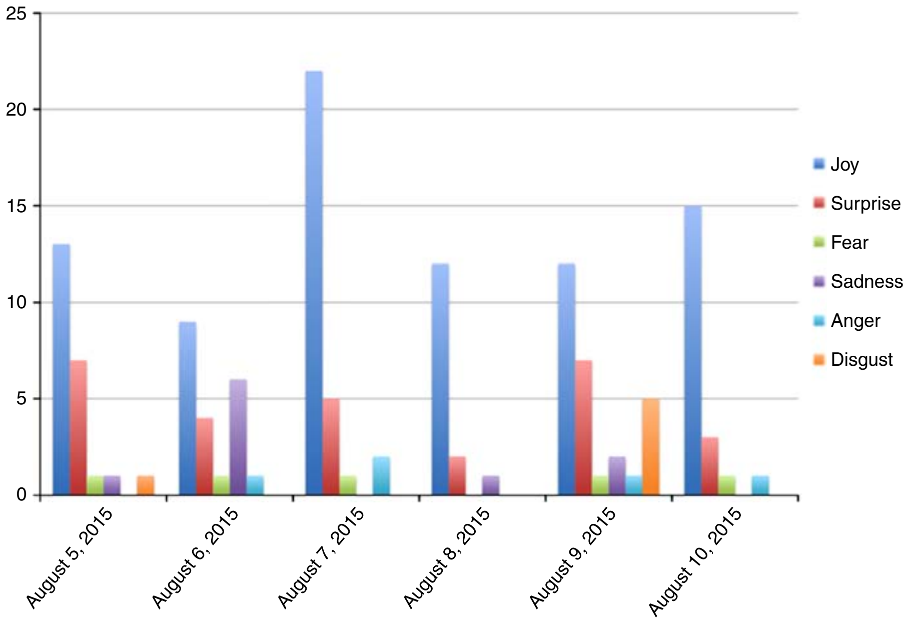
The polarity chart shows that there is a peak of positive sentiments on August 7, given by posts like:

- Fri Aug 07 20:09:13 and Ted Cruz so much +0000 2015 - I was very impressed with these two! http:b'RT @JoeTheMailman: #GOPdebate//t.co/gDEDcJ89Ne'I appreciate Ben Carson



Note: Percentages on a sample of 300 tweets retrieved

Figure 11. Polarity chart of tweets regarding Ben Carson



Note: Percentages on a sample of 300 tweets retrieved

Figure 12.
Emotions chart of tweets regarding Ben Carson

- Fri Aug 07 20:06:12 during last nights #GOPDebate @RealBenCarson http:+0000 2015 - b'RT @ThePatriot143: Dr Ben Carson was very impressive//t.co/fyFdNokrHH'
- Fri Aug 07 20:05:46 have life first "-Dr Ben Carson '+0000 2015 - b' "You can´t have liberty and the pursuit of happiness if you don't

For what concerns the "joy" emotion we have 14 tweets like:

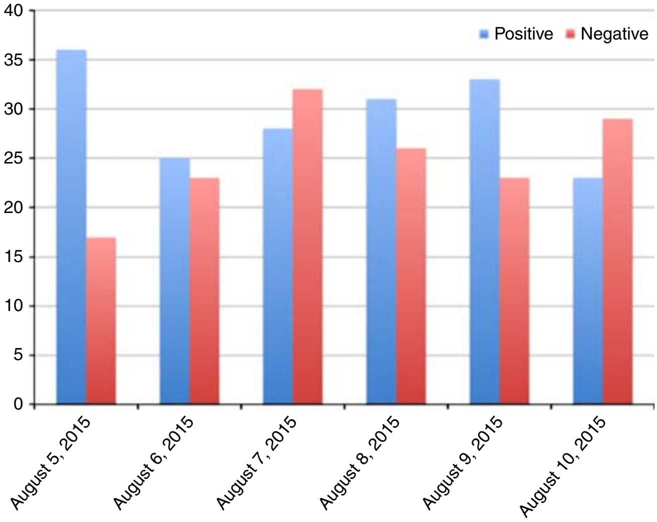
- Fri Aug 07 20:01:01them who they are […] it+s time for us to move beyond that." #GOPDebate http:´ 0000 2015 - b'RT @ABC: Dr Ben Carson on race: "The skin doesn//t.co/o'´t make

Other tweets automatically labeled as "joy" by the system are:

- Fri Aug 07 20:00:47 wow i love this guy #drbencarson'+0000 2015 - b'dr ben carson is such a breath of fresh air in american politics.
- Fri Aug 07 19:55:53 date […] but Ben Carson sure seems to have a good head on his shoulders"+0000 2015 - b"I've liked Donald trump from day one as a presidential candi-
- Fri Aug 07 19:51:15 +0000 2015 - b'RT @Talkmaster: I like Ben Carson a lot more after this debate' On August 9th there is a peak of "surprise" given by some tweets like:
- Sun Aug 09 15:37:57 Only real surprise for me was Ben Carson's fabulous sense of humor. What a grea"+0000 2015 - b"RT @LindaSuhler: Not sure any minds were changed tonight.

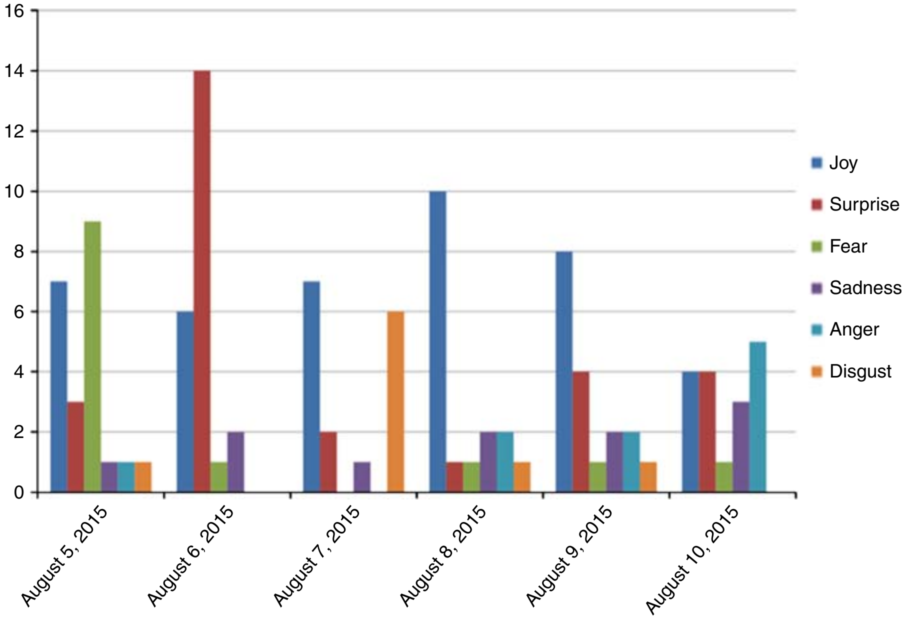
- Sun Aug 09 15:40:48 +0000 2015 - b'RT @kirstiealley: Dr Ben Carson is sorta amazing.'

(1) Scott Walker: for the candidate "Scott Walker" the charts related to the sentiment and the emotions detected among 300 retrieved tweet are reported in Figures 13 and 14.



Note: Percentages on a sample of 300 tweets retrieved

Figure 13. Polarity chart of tweets regarding Scott Walker



Note: Percentages on a sample of 300 tweets retrieved

Figure 14. Emotions chart of tweets regarding Scott Walker

From the chart illustrated in Figure 13 we can observe that on August 5th there is a peak of posts automatically labeled by the system as "positive." As an example, we report some of them as follows:

- Wed Aug 05 22:03:21 getting ready to spend millions on television and radio ads in Iowa and other […]"+0000 2015 - b"A super PAC supporting Scott Walker's run for president is
- Wed Aug 05 21:44:45 walker is for some hot, hot ham"+0000 2015 - b"@swarthyvillain i don't think anyone can be hornier than scott
- Wed Aug 05 22:17:42 most underhanded politician in America. And thats saying something.'+0000 2015 - b'Scott Walker wins my Tom Delay award for sneakiest, slimiest, Wed Aug 05 22:09:21 http://t.co/5kVMomXKej'+0000 2015 - b'Facing rising Trump, Walker team fortifies in Iowa

Another interesting day is August 7, where we can observe a peak of posts labeled as "negative," some of which are reported as follows:

- Fri Aug 07 20:12:56 His Anti-Choice Extremism http:+0000 2015 -//t.cob'iNewsReport: Scott Walkers False Claim That America Shares/g5TuGlSDOo #Election2016'
- Fri Aug 07 20:11:25 InTheseTimesMag'+0000 2015 - b'Scott Walker Is a Dictator-In-Waiting http://t.co/OrhUGfptz9
- Fri Aug 07 20:09:05 investigation in 2011 http:+0000 2015 -//t.co/q8YB3W9lbj http: b'RT @thedailybeast: Scott Walker was named in a criminal//t.co/DIAKOUsnqt'
- Fri Aug 07 20:05:01dying" Die, already, won+0000 2015 -´t you?" #scottwalker #abortion'b'Scott Walker tells pregnant woman who can´t bear a child w/o

We report also some of the tweets automatically labeled as "positive" on the same day:

- Fri Aug 07 20:19:41 Hillary Clinton's e-mail scandal: http:+0000 2015 - b"RT @Slate: Scott Walker made a legitimately funny joke about//t.co/eIsOOCStnV #GOPDebate http://t.co"
- Fri Aug 07 20:05:53 America" - Scott Walker'+0000 2015 - b'RT @Brentweets: "I saved Wisconsin, cheese is back, I can save
- Fri Aug 07 20:05:28 DougCollins as my#Walker16 Georgia campaign chair. http: +0000 2015 - b"RT @ScottWalker: I'm proud to announce and welcome @Rep//t.co/0Z0Ee3pzNZ - SW"-
- Fri Aug 07 20:13:31 He's worried about Republicans who think Trump is too tolerant. https:+0000 2015 - b"RT @LOLGOP: Scott Walker isn't worried about Americans.//t.co/Qtbh9h"

For what concerns the emotions, there is a huge hike on August 6 for the emotion "surprise," a maximum on August 8 of joy and fear on August 5. The main reason of the peak of "fear" on August 5 is a set of 26 tweets labeled as "fear" like this:

- Wed Aug 05 21:58:20 this week's issue +0000 2015 - b"RT @thenation: Meet the absolutely ruthless Scott Walker in ×97totally free for all to read: http://t.co/o5PUED2wEY http://t"

On the same day there is also a relevant percentage of "joy" due to a set of five joy tweets like this:

- Wed Aug 05 22:09:07 team fortifies in Iowa http:+0000 2015 -//t.co/M01BJ83phe - good reporting from @tomlobianco"b"RT @betsy klein: With Donald Trump at door, Scott Walker's

We can observe also four tweets labeled as "surprise" like this:

- Wed Aug 05 22:14:18 Donald Trump. Maybe he could get a selfie, or another campaign contribution. http:+0000 2015 - b'RT @NicholsUprising: Scott Walker will be right next to/'

  The peak of "surprise" on August 6 is due to 28 messages like:

- Thu Aug 06 20:14:45 can get a new sports stadium+0000 2015 -/arena built wb"RT @BillSimmons: I can't believe that, in 2015, any billionaire/ taxpayer money. http://t.co/D5a"

Another kind of message automatically labeled as "surprise" is:

- Thu Aug 06 21:02:00 surprised? […] Anyone? http:+0000 2015 -//t.co/SYC2Hi3Xm6 http: b'Oh look. @ScottWalker wins the#KochBros straw poll. Anyone//t.co/iygqFtw5Qc'

The interesting thing in this case is that the percentage level of "positive" sentiment for Donald Trump overcomes that one of the other candidates, while the analysis of the emotions would require a deeper and extensive inspection.

6. Conclusions

The paper has focused on the employment of social media as indicators of public opinions of the real-world phenomena. With the recent development and widespread use of Google Trends, we have a promising tool for automatically analyzing and "sense" social media looking for sentiment orientations and emotions arising about very specific topics, which are at the moment identified as "hot," due to the high search volume of specific keywords in queries. According to these considerations, we have developed and illustrated a framework that, triggered by "hot queries" coming from Google Trends, tries to measure both the sentiment and the emotions expressed in posts broadcasted on the Twitter social network. We have preferred Twitter since it is one of the most representative microblogging platforms, which allows users to upload new messages more frequently and instantly with respect to other social networks. We have tested the platform on several hot trend stories, and we have reported in this paper three case studies ranging from consumer electronics, public health and politics, topics which are in context with the literature illustrated in the first part of the paper. The results clearly show that, even if more studies and work has to be done, computational intelligence can help in inferring real-world phenomena through social media sentiments and emotions. The automatic translation process has been considered in our system in order to make the approach more general, however, we decided in a second phase to consider only tweets written in English language in order to limit as much as possible the bias, with the goal of substituting, in future works, the automatic translation of tweets with a set of more effective multi-lingual sentiment and emotion classifiers.

Notes

1. When we talk about consumers, we refer to the broader meaning of the term. For us examples of consumers are: a buyer who purchases products on an e-commerce website; a public administrator who has to decide on local health policy; an epidemiologist interested to know the feeling of the public on the propagation of a virus; an economist interested in the dynamics of the labor market; and so forth. These examples show how we opt in favor of an operational definition, aimed at what we are going to introduce later.
2. The keywords users enter for their search.
3. www.google.com/trends/
4. www.wordstream.com/blog/ws/2011/05/25/keywords-vs-search-queries
5. http://techcrunch.com/2008/10/09/some-big-sites-are-using-google-trends-to-direct-editorial
6. https://en.wikipedia.org/wiki/Computational-epistemology

7. The whole interview, released on March 2008, is available at: www.ft.com/cms/s/0/d386202c-f3c3-11dc-b6bc0000779fd2ac.html♯axzz3haIRECAh
8. CoreLogic-CoreLogic Reports 57,000 Completed Foreclosures in September – September 2012 http://multivu.prnewswire.com/mnr/corelogic/56990/
9. https://business.twitter.com
10. http://sentiwordnet.isti.cnr.it
11. www.cs.uic.edu/ liub/FBS/ sentiment-analysis.html
12. http://mpqa.cs.pitt.edu/ opinionfinder
13. The authors claim that they used all the polls published in MetaPolls (http://metapolls.net); further resources included www.wahlrecht.de for Germany, www.3comma14.gr for Greece, and polls from Ipsos (www.ipsos-ned- erland.nl), TNS Nipo (www.tns-nipo.com), and Peil (www.noties.nl/ peil.nl) for the Netherlands.
14. Twitter developers: Streaming API methods, resources available at: https://dev.twitter.com/docs/streaming-api/methods
15. Each JavaScript Object Notation contains one tweet.
16. www.nltk.org
17. www.meaningcloud.com

References

Ahn, H.-I and Spangler, W.S (2014), “Sales prediction with social media analysis”, *Global Conference (SRII), 2014 Annual SRII, IEEE*, *San Jose*, *April 23-25*.

Asur, S. and Huberman, B. (2010), “Predicting the future with social media”, *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Vol. 1, IEEE*, *Toronto*, *August 31-September 3*.

Burnap, P., Gibson, R., Sloan, L., Southern, R. and Williams, M. (2015), “140 characters to victory?: using Twitter to predict the UK 2015 general election”.

Cao, G., Wang, S., Hwang, M., Padmanabhan, A., Zhang, Z. and Soltani, K. (2015), “A scalable framework for spatiotemporal analysis of location-based social media data”, *Computers, Environment and Urban Systems*, Vol. 51, pp. 70-82.

Ceron, A., Curini, L., Iacus, S.M. and Porro, G. (2014), “Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens’ political preferences with an application to Italy and France”, *New Media & Society*, Vol. 16 No. 2, pp. 340-358.

D’Avanzo, E. and Pilato, G. (2015), “Mining social network users opinions’ to aid buyers’ shopping decisions”, *Computers in Human Behavior Part B*, Vol. 51, pp. 1284-1294.

Dhar, R. and Simonson, I. (2003), “The effect of forced choice on choice”, *Journal of Marketing Research*, Vol. 40 No. 2, pp. 146-160.

Do, H.J. and Choi, H.-J. (2015), “Korean Twitter emotion classification using automatically built emotion lexicons and fine-grained features”, *29th Pacific Asia Conference on Language, Information and Computation*, *Shanghai*, *October 30-November 1*, pp. 142-150.

Dzielinski, M. (2012), “Measuring economic uncertainty and its impact on the stock market”, *Finance Research Letters*, Vol. 9 No. 3, pp. 167-175.

Eichstaedt, J.C., Schwartz, H.A., Kern, M.L., Park, G., Labarthe, D.R., Merchant, R.M., Jha, S., Agrawal, M., Dziurzynski, L.A., Sap, M., Weeg, C., Larson, E.E., Lyle, H., Ungar, L.H. and Seligman, M.E.P. (2015), “Psychological language on Twitter predicts county-level heart disease mortality”, *Psychological Science*, Vol. 26 No. 2, pp. 159-169.

Ekman, P. (1992), “An argument for basic emotions”, *Cognition and Emotion*, Vol. 6 Nos 3-4, pp. 169-200.

Gao, Y., Wang, F., Luan, H. and Chua, T. (2014), “Brand data gathering from live social media streams”, *Proceedings of International Conference on Multimedia Retrieval, ACM*, *Glasgow*, *April 1-4*.

Gayo-Avello, D.A. (2013), "Meta-analysis of state-of-the-art electoral prediction from Twitter data. arXiv preprint".

Ginsberg, J., Mohebbi1, M.H., Patel, R.S., Brammer, L., Smolinski, M.S. and Brilliant, L. (2009), "Detecting influenza epidemics using search engine query data", *Nature*, Vol. 457 No. 7232, pp. 1012-1014.

Goel, S., Hofman, J., Lahaie, S., Pennock, D.M. and Watts, D.J. (2010), "Predicting consumer behavior with web search", *Proceedings of the National Academy of Sciences*, Vol. 107 No. 41, pp. 17486-17490.

Karagiorgou, S., Pfoser, D. and Skoutas, D. (2014), "Geosemantic network-of-interest construction using social media data", in Duckham, M., Pebesma, E., Stewart, K. and Frank, A.U. (Eds), *Geographic Information Science*, Springer International Publishing, Berlin, pp. 109-125.

Kitchin, R. (2014), "Big data, new epistemologies and paradigm shifts", *Big Data & Society*, Vol. 1 No. 1, pp. 1-12.

Lagre, P., Cautis, B. and Vahabi, H. (2015), "A network-aware approach for searching as-you-type in social media", *Conference on Information and Knowledge Management*, *Melbourne*, *October 19-23*.

Leginus, M., Zhai, C.X. and Dolog, P. (2015), "Beomap: ad hoc topic maps for enhanced exploration of social media data", in Cimiano, P., Frasincar, F., Houben, G.-J. and Schwabe, D. (Eds), *Engineering the Web in the Big Data Era*, Springer International Publishing, Berlin, pp. 200-218.

Levenberg, A., Pulman, S. and Mollanen, K. (2014), "Predicting economic indicators from web text using sentiment composition", *International Journal of Computer and Communication Engineering*, Vol. 3 No. 2, pp. 109-119.

Liu, B. (2005), *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*, Cambridge University Press, Cambridge.

Liu, Y., Sui, Z., Kang, C. and Gao, Y. (2014), "Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data", *PloS One*, Vol. 9 No. 1, pp. 1-11.

Loeb, S., Bayne, C.E., Frey, C., Davies, B.J., Averch, T.D. and Woo, H.H. (2014), "Use of social media in urology: data from the American Urological Association (AUA)", *BJU International*, Vol. 113 No. 6, pp. 993-998.

Middleton, S.E., Middleton, L. and Stefano, M. (2014), "Real-time crisis mapping of natural disasters using social media", *IEEE Intelligent Systems*, Vol. 29 No. 2, pp. 9-17.

Mohammad, S.M., Salameh, M. and Kiritchenko, S. (2016), "How translation alters sentiment", *Journal of Artificial Intelligence Research*, Vol. 55 No. 1, pp. 95-130.

Moon, S., Bae, S. and Kim, S. (2014), "Predicting the near-weekend ticket sales using web-based external factors and box-O ce data", *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 02, IEEE Computer Society*, *Warsaw*, *August 11-14*.

Ruths, D. and Pfeffer, J. (2014), "Social media for large studies of behavior", *Science*, Vol. 346 No. 6213, pp. 1063-1064.

Sangiorgi, P., Augello, A. and Pilato, G. (2013), "An unsupervised data-driven cross-lingual method for building high precision sentiment lexicons", *2013 IEEE Seventh International Conference on Semantic Computing (ICSC), IEEE, September*, Vol. 22 No. 2, pp. 189-214.

Schivinski, B. and Dariusz, D. (2014), "The effect of social media communication on consumer perceptions of brands", *Journal of Marketing Communications*, Vol. 22 No. 2, pp. 189-214.

Scott, S.L. and Varian, H. (2014), "Bayesian variable selection for nowcasting economic time series", in Goldfarb, A., Greenstein, S.M. and Tucker, C. (Eds), *Economic Analysis of the Digital Economy*, University of Chicago Press, Chicago, IL, pp. 119-135.

Song, G., Cheon, Y., Lee, K., Lim, H., Chung, K. and Rim, H. (2014), "Multiple categorizations of products: cognitive modeling of customers through social media data mining", *Personal and Ubiquitous Computing*, Vol. 18 No. 6, pp. 1387-1403.

Spangler, S., Wilkins, A.D., Bachman, B.J., Nagarajan, M., Dayaram, T., Haas, P., Regenbogen, S., Pickering, C.R., Comer, A., Myers, J.N., Stanoi, I., Kato, L., Lelescu, A., Labrie, J.J., Parikh, N., Lisewski, A.M., Donehower, L., Chen, Y. and Lichtarge, O. (2014), "Automated hypothesis generation based on mining scientific literature", *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM*, *New York, NY*, *August 24-27*.

Strapparava, C., Valitutti, A. and Stock, O. (2006), "The affective weight of the lexicon", *5th International Conference on Language Resources and Evaluation (LREC 2006)*, *Genoa*, pp. 474-481.

Tang, J., Chang, Y. and Liu, H. (2014), "Mining social media with social theories: a survey", *ACM SIGKDD Explorations Newsletter*, Vol. 15 No. 2, pp. 20-29.

Terrana, D. and Pilato, G. (2013), "Detection, clustering and tracking of life cycle events on Twitter using electric fields analogy", *IEEE Seventh International Conference on Semantic Computing (ICSC)*, *September 16-18*, pp. 220-227.

Tripathi, A., Hossain, S., Singh, V.K. and Atrey, P.K. (2015), "Assessing personality using demographic information from social media data", *Proceedings of the 2015 International Conference on Social Media and Society, ACM.*

Tsakalidis, A., Papadopoulos, S., Cristea, A.I. and Kompatsiaris, Y. (2015), "Predicting elections for multiple countries using Twitter and polls", *IEEE Intelligent Systems*, Vol. 30 No. 2, pp. 10-17.

Varian, H.R. (2014), "Big data: new tricks for econometrics", *The Journal of Economic Perspectives*, Vol. 3 No. 2, pp. 3-27.

Webb, G.K. (2009), "Internet search statistics as a source of business intelligence: searches on foreclosure as an estimate of actual home foreclosures", *Issues in Information Systems*, Vol. 10 No. 2, pp. 82-87.

White, C. (2011), "Using big data for smarter decision making", BI research.

Williamson, J. (2008), "The philosophy of science and its relation to machine learning", in Gaber, M.M. (Ed.), *Scientific Data Mining and Knowledge Discovery: Principles and Foundations*, Springer, Berlin, pp. 77-89.

Wilson, T., Wiebe, J. and Hoffmann, P. (2005), "Recognizing contextual polarity in phrase-level sentiment analysis", *Proceedings of HLT-EMNLP-2005*, *Vancouver*, *October 6-8*.

Xia, C., Schwartz, R., Xie, K., Krebs, A., Langdon, A., Ting, J. and Naaman, M. (2014), "CityBeat: real-time social media visualization of hyper-local city data", *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee*, *Seoul*, *April 7-11*.

Xu, B., Yuan, T.C.W., Fussell, S.R. and Cosley, D. (2014), "SoBot: facilitating conversation using social media data and a social agent", *Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing, ACM*, *Baltimore*, *February 15-19*.

Yun, J.-S., Park, J.T., Hwang, H.S. and Moon, I.Y. (2015), "Customized efficient collection of big data for advertising services".

Zhang, Z., Wang, S., Cao, G., Padmanabhan, A. and Wu, K. (2014), "A scalable approach to extracting mobility patterns from social media data", *2014 22nd International Conference on Geoinformatics (GeoInformatics), IEEE*, *Kaohsiung*, *June 25-27*.

Zhu, J., Wang, X., Qin, J. and Wu, L. (2012), "Assessing public opinion trends based on user search queries: validity, reliability, and practicality", *Annual Conference of the World Associate for Public Opinion Research*, *Hong Kong*.

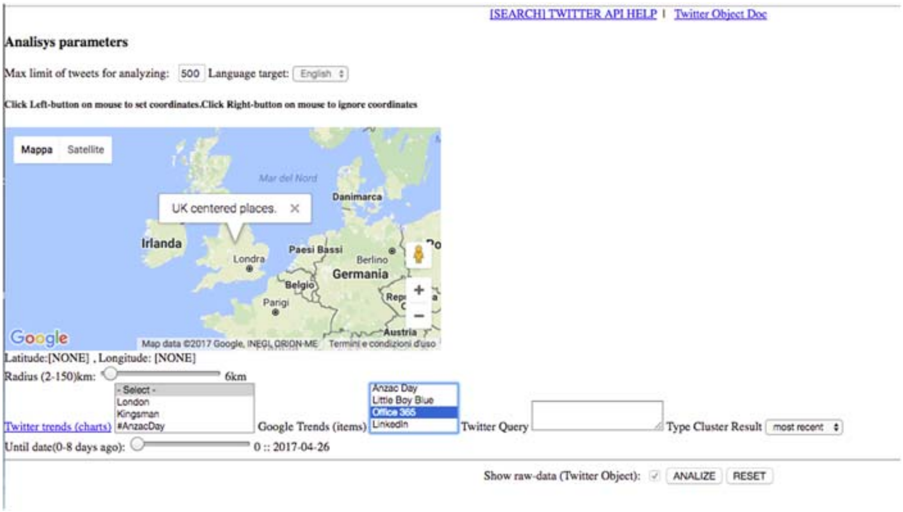
Further reading

Castillo, C, El-Haddad, M, Pfeffer, J and Stempeck, M (2014), "Characterizing the life cycle of online news stories using social media reactions", *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing, ACM*, *Baltimore*, *February 15-19*.

Appendix. A case study

In the following, we illustrate a step-by step example. In particular, from our dashboard, we have detected as one of the most the hot topic from Google Trends that one regarding "Office 365," so we decided to analyze that specific topic by retrieving 500 tweets from the area of London (within a radius of 150 Km), see Figure A1. We obtained the results illustrated in Figures A2 and A3. In particular, some of the retrieved tweets are:

- Enjoy a fully installed Odon #UK https://t.co/xFyLvo4QGsce experience across devices #O ce365 #Microsoft #Vodanile #Lon-
- Microsoft announces major changes to Oce 365 system requirements https://t.co/HhtDw5g6YK
- RT @cogmotive: Working smarter, not harder, with #ArtificialIntelligence and #Ohttps://t.co/HLC4zO1w1z #AI #O365 https://t.co/ur5?ce365:
- Microsoft will block Oce 2016 users from accessing Oce 365 https://t.co/hSRnNdS9GZ
- RT @mspoweruser: You will soon be able to purchase Odows Store - https://t.co/JC6Z7oeSyY https://t.co/p? ce 365 subscriptions from the Win-
- Our April #Ocrosoft? https://ce365 Update is out, with new features & important announcements from Mi-t.co/8GOTjNu3mo



[SEARCH] TWITTER API HELP | Twitter Object Doc

Analisys parameters

Max limit of tweets for analyzing: 500 Language target: English

Click Left-button on mouse to set coordinates.Click Right-button on mouse to ignore coordinates

Latitude:[NONE] , Longitude: [NONE]

Radius (2-150)km: 6km

Twitter trends (charts): - Select - London Kingsman #AnzacDay

Google Trends (items): Anzac Day Little Boy Blue Office 365 LinkedIn

Twitter Query

Type Cluster Result: most recent

Until date(0-8 days ago): 0 :: 2017-04-26

Show raw-data (Twitter Object): ✓

Figure A1.
The screenshot of the decision support tool

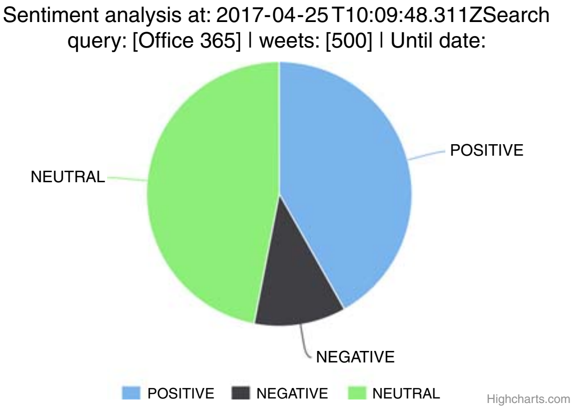


Figure A2.
The results obtained for "sentiment" regarding the "O ce 365" Topic

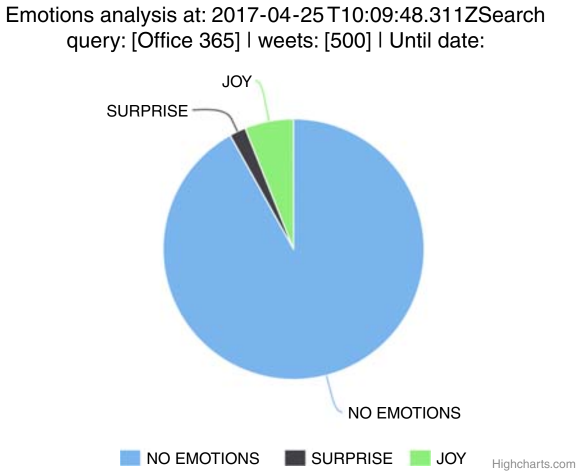


Figure A3.
The results obtained for "emotions" regarding the "Office 365" Topic

Furthermore, we see that the most common adjectives or adverbs in tweets are "new" (20.87 percent), "able" (8.27 percent), "soon" (7.87 percent), "major" (5.91 percent), "easily" (4.72 percent), "free" (3.54 percent), "available" (2.76 percent).

According to these results, being them "positive" (41 percent), and triggering "joy" (7 percent) and "surprise" (2 percent), the user of the system will probably have a proclivity for buying in the future that product, or at least on being up-to-date about the development of the product.

About the authors

Ernesto D'Avanzo teaches Philosophy of Science and Decision Making for Degree in Communication Sciences where he also teaches Data, Text and Social Analytics for Master courses in Communication Sciences. He investigates decision models and decision support systems that make the use of artificial intelligence methods and techniques (e.g. machine learning and natural language processing) and mathematical models (path analysis, confirmatory factor analysis, latent growth modeling). Current research interests and projects concern, in particular, biomedical and healthcare decision making, consumer/buyer's behavior modeling, and socio-political analysis and decision making. He serves as an Associate Editor of the *International Journal on Semantic Web and Information Systems*. Ernesto D'Avanzo is the corresponding author and can be contacted at: edavanzo@unisa.it

Giovanni Pilato received his "cum laude" Laurea Degree in Electronic Engineering and the PhD Degree in Computer Science from the University of Palermo, Italy, in 1997 and 2001, respectively. He is currently a Staff Research Scientist at the ICAR-CNR (Italian National Research Council) branch of Palermo, Italy. He is also a Lecturer at the Mathematics and Computer Science Department of the University of Palermo, Italy. His research interests include geometric techniques for knowledge representation, web data mining, social sensing and natural language processing.

Miltiadis D. Lytras is a Research Professor in the American College of Greece, with a research focus on Robotics, Cognitive Computing, information systems, technology enabled innovation, social networks, computers in human behavior, semantic web, knowledge management and technology enhanced learning, with more than 150 publications in these areas. He serves as the Editor in Chief of the *International Journal on Semantic Web and Information Systems* and the Editor in Chief, *International Journal on Knowledge Society Research*. He is a distinguished Scholar in King Abdulaziz università.