Undefined 0 (0) 1
IOS Press

What'chu talkin' about, Willis?

A consumer's look on Facebook and Twitter – What do people read and where?

Thomas Steiner [a], Ruben Verborgh [b], Raphaël Troncy [c], Giuseppe Rizzo [c], Rik Van de Walle [b], Joaquim Gabarro [d], Arnaud Brousseau [a,*]

[a] *Google Germany GmbH, ABC-Str. 19, 20354 Hamburg, Germany,*
E-mail: tomac@google.com, arnaud.brousseau@gmail.com
[b] *Ghent University – IBBT, ELIS, Multimedia Lab, Gaston Crommenlaan 8/201, 9050 Ghent, Belgium,*
E-mail: ruben.verborgh@ugent.be, rik.vandewalle@ugent.be
[c] *EURECOM, Sophia Antipolis, France*
E-mail: raphael.troncy@eurecom.fr, giuseppe.rizzo@eurecom.fr
[d] *Universitat Politècnica de Catalunya, Department LSI, 08034 Barcelona, Spain,*
E-mail: gabarro@lsi.upc.edu

Abstract. With the ever-growing influence of social networks, social media mining becomes more and more important as a source for responses to all sorts of questions. *"Do people like product X?"*; *"What do people think of a new law proposal Y?"*; *"Will candidate A or candidate B win the elections?"*. These are just some sample questions where social networks can substantially contribute to answers. In this paper, we propose a paradigm shift in order to find responses. Where traditional social media mining focuses exclusively on the producer side of microposts, we focus on the consumer side, that is, on the readers of microposts. Traditional social media mining retrieves its data through official Application Programming Interfaces (APIs). In contrast, our approach works through accessing its data via browser extensions directly from the social network users' timeline when they visit their social network of choice via a Web browser. In comparison to social data retrieved via APIs, the social data retrieved via our approach is more sparse, however, we argue in the paper that it is of higher quality. We have implemented browser extensions for the popular social networks Facebook and Twitter. These extensions perform named entity disambiguation on microposts and, via Web analytics software, enabled us to collect social data over the course of six months. In the first part of the paper, we present global statistics and a comparison of what topics people are interested in on the two examined social networks. In the second part, using concrete examples from recent history, we show how additional data gathered through Web analytics software can be used to get fine-grained information on geolocations of centers of interest. This allows for interesting new kinds of questions to be addressed. *"Does an event X cause more reader interest in country A than in country B?"*; *"Which continent cares most about a catastrophe Y?"*; *"Do people in city Z read about product P?"*. Finally, as our approach allows for cross-network *ambiguity-free* social media mining, we can even propose answers for a question like the following: *"Is my brand B read more about in region R on social network A, or social network B?"*. We see our approach not as a replacement of traditional social media mining, but more as an additional perspective that makes sense in certain scenarios, some of which we present in this paper.

Keywords: Social media mining, named entity disambiguation, Web analytics

[*] A. Brousseau was an intern at Google Germany GmbH at time of writing.

1. Introduction

1.1. On Traditional Social Media Mining

In recent years, social media mining has become an essential tool for marketers, traders, and researchers. The information people share publicly via so-called *microposts* on social networks harbor tremendous amounts of valuable social data. Forbes has called the social graph *crude oil* in a recent blog post [57]:

> "The point is, crude oil is crude. It is an unrefined and complex natural resource containing many riches. It takes time to figure out what to do with a new crude resource."

1.2. Social Network Data Access via APIs

Social networks today are very much seen as "walled gardens", excellently illustrated by a cartoon by David Simonds (Figure 1). This network isolatedness reflects on how traditional social media mining is done nowadays. Common literature typically either focuses on just one network (*e.g.*, [43]), or treats the different networks separately (*e.g.*, [44]). Traditional social media mining happens (i) based on either term-based search APIs, and/or (ii) based on so-called "fire hose" near-realtime streaming APIs, which are both provided by the social networks themselves. The main difference between (i) and (ii) is that, in the prior case, terms like the name of a brand or company are *proactively* searched for, whereas in the latter case the social media mining system *reactively* acts upon the occurrence of such terms.

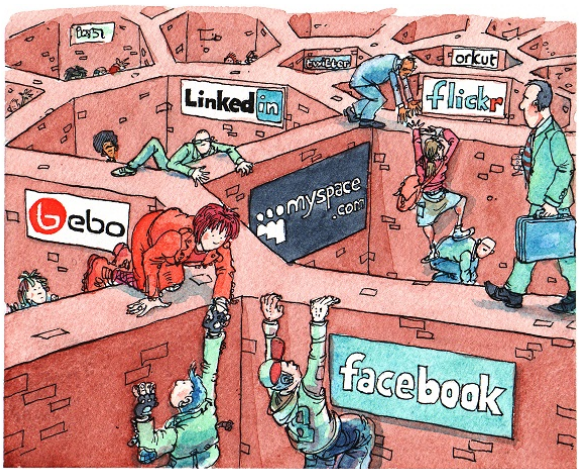


Fig. 1. David Simonds illustrates social networks as walled gardens due to their (by design) lock-in effects [13].

1.3. Selection of Social Networks and APIs

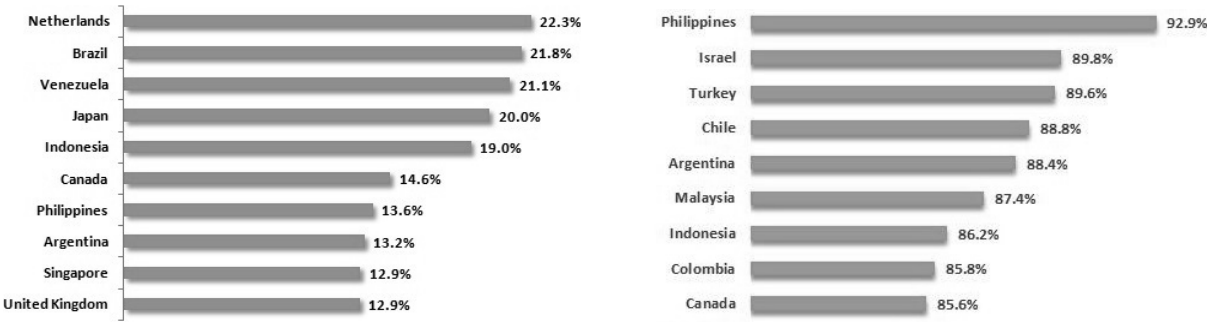
In this paper, we consider the popular social networks Facebook [15] and Twitter [54], currently the two globally most important social networks [10,11]. Figure 2 shows the percentages of the online population of several countries and their usage of Facebook and Twitter. Traditionally, Twitter is very permissive with its API, as since the beginning of the platform, API-based Twitter clients play a strategic role for the company. Twitter provides developers with the Twitter Streaming API [56], which allows for high-throughput near-realtime access to various subsets of public and protected Twitter data, at a coverage rate of 1% ("sprinkler"), 10% ("garden hose"), or 100% ("fire hose") of all Twitter traffic. Facebook has no such public "fire hose" streaming API, but supports near-realtime updates via its Graph API [16] to enable applications to subscribe to a limited set of changes in data. Whenever such a change occurs, Facebook notifies subscribers with a list of changes. The obvious issue here is that, in order to get a Twitter Streaming API-like experience, one has to subscribe to an impossibly high number of users. This imbalance in data availability via the respective APIs has an impact on academic publications on social network mining. While at the World Wide Web Conference 2011 (WWW2011) alone, three Twitter papers based on the Twitter API were published [30,42,59], publications on Facebook typically focus on privacy issues (*e.g.*, [28]), or Facebook's sociological impact (*e.g.*, [14]), without making use of the Facebook API.

1.4. Positioning of our Work

What, to the best of our knowledge, all publications so far have in common is their focus on the author side: it is very well researched what people *produce* on social networks (especially Twitter), whom they follow or unfollow and why, what they tag, whom they put in what list, group, or circle, etc. However, few to no focus has been put on what people *consume* – or at least no such study is publicly available. This is especially true *across* social networks. As far as we can tell, no study has compared *reader* behavior on *different* social networks in parallel before.

1.5. Overview of our Data Retrieval and Enrichment Processes

In this work, we thus compare topics people read about on Facebook and on Twitter, and classify those



(a) Twitter covers at most one in five people per country [10].

(b) Facebook covers the majority of people in several countries [11].

Fig. 2. A significant percentage of the online population participates in social networks.

topics in order to provide an overall comparison. We therefore have implemented two similar browser extensions: Twitter Swarm NLP[1] and Facebook Swarm NLP[2]. On the one hand, these extensions enrich the user experience on the two social networks Facebook and Twitter, and on the other hand, they determine the topics that people see and read about on their timelines by means of named entity disambiguation. We use a definition of *named entity* that was coined by Grishman *et al.* as *an information unit described by the name of a person or an organization, a location, a brand, a product, a numeric expression including time, date, money and percent found in a sentence* [20]. We combine this knowledge with additional anonymous data that we obtain about social network users through Web analytics software.

1.6. Focus on Desktop Browser Versions

The two implemented browser extensions require a desktop browser in order to work. Whenever a user visits a social networking site, in the concrete case Facebook or Twitter, the particular extension gets activated. By focusing exclusively on content people see when directly navigating to the desktop versions of either `twitter.com` or `facebook.com` – therefore on purpose neglecting all activity via applications on *both* desktop and mobile devices – we assume people indeed read that content. This is justified by each site's requirement to manually click a link "n new stories" (Facebook) or "n new tweets" (Twitter) for new content to appear, rather than auto-updating the timeline. Other approaches to determine whether a micropost has been read are limited to microposts with contained Web links and checking whether clicks on those links have occurred. However, automatic crawling and indexing of links adds hard to detect noise. We therefore argue that our approach has a higher precision, at the cost of lower recall.

1.7. Paper Objective and Structure

We outline our paper objectives and non-objectives explicitly, where each objective has a corresponding non-objective in the lists below. In this paper, we **will**:

- perform analyses based on disambiguated named entities;
- perform analyses based on IP-address-based reader location detection;
- work with a manageable amount of microposts read by a random population of social network users;
- focus on the micropost reader side.

On the contrary, we **will not**:

- perform analyses based on hashtags, term frequencies, or trends;
- perform analyses based on natively geotagged microposts;
- work with huge amounts of microposts from "fire hose" APIs;
- focus on the micropost author side,

which is why we strive for a paradigm shift that promises new insights for tasks like brand analysis, opinion research, but also sociological questions.

The remainder of this paper is structured as follows. Section 2 focuses on structuring and consolidating unstructured textual micropost data, and introduces browser extensions and Web analytics software.

[1] Twitter Swarm NLP: `http://bit.ly/twitterswarmnlp`
[2] Facebook Swarm NLP: `http://bit.ly/facebookswarmnlp`

Section 3 explains our experiment setup and gives an overview of user demographics. Section 4 starts with a presentation of raw statistic, then provides a ranking of named entities and gives a categorization of the RDF types of the named entities. The Section ends with some scenarios from recent history where we show how our approach can be used to reveal new insights that would not be possible with traditional social media mining. We report on related work in Section 5, and give an outlook on future work in Section 6. We close the paper with a conclusion in Section 7.

2. Implementation

We have implemented two browser extensions to cover the social networking sites Facebook and Twitter. These extensions were released for free on a Web store for browser extensions with the following description (slightly adapted):

> This extension performs Named Entity Extraction (NEE) on the microposts you read and write on {Twitter, Facebook}. If you write: "Had froyo for breakfast.", a named entity would be "froyo". In the sense of Linked Data, we identify such named entities via a URI, for example `http://dbpedia.org/resource/Frozen_yogurt` in the concrete case. You can see the extracted entities highlighted in each status message *(see Figure 3)*. In addition to that, the extracted entities are then reported to a shared Web analytics account via event tracking code that allows us to build a ranking of the most-talked-about entities.

The extensions were released for the Google Chrome browser, the Web Analytics software that we used was Google Analytics. Micropost texts are sent to a server that performs named entity extraction. Afterwards, the original text is discarded and only the extracted entities remain on the server for analysis.

2.1. Structuring Unstructured Data

A priori, microposts are unstructured textual data. We apply so-called Linked Data rules in order to convert this unstructured data into structured data. In a first step, the process consists of named entity detection via Natural Language Processing (NLP), and in a second step, named entity disambiguation. Sir Tim Berners-Lee has introduced Linked Data in a W3C Design Issue [6], where he defines the four rules for Linked Data as follows:

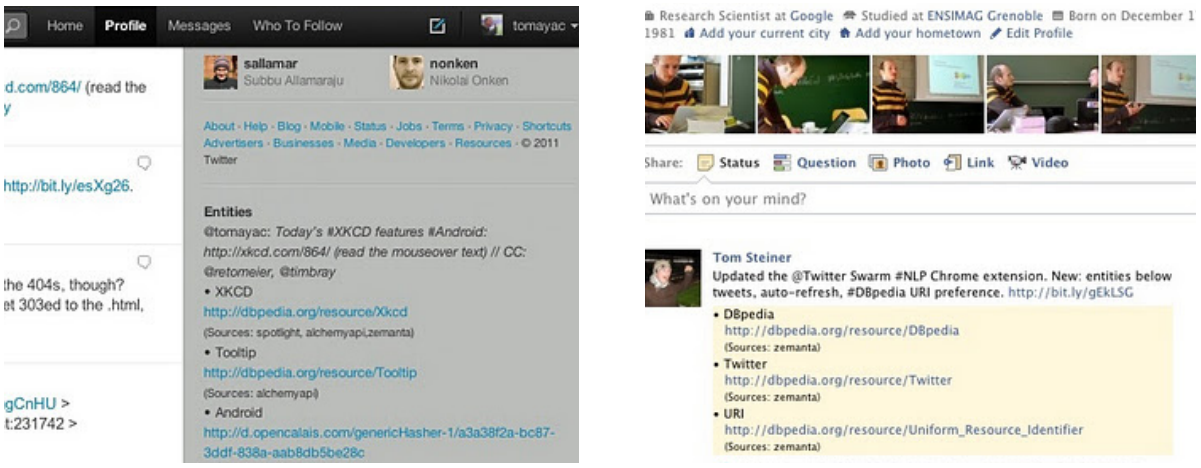
1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL).
4. Include links to other URIs, so that they can discover more things.

In order to represent extracted named entities from microposts in an unambiguous way, we apply the first and the second Linked Data principle by representing disambiguated named entities with HTTP URIs. We outsource this task to third party named entity disambiguation Web services (APIs), namely OpenCalais [36], Zemanta [60], DBpedia Spotlight [32], and AlchemyAPI [1]. These APIs take a text fragment as an input, perform named entity extraction and disambiguation on it, and then link the extracted named entities back into the Linking Open Data (LOD) cloud [12]. We use these APIs in parallel, and, by combining their results [40], aim at the emergence effect in the sense of Aristotle: *"[. . .] the totality is not, as it were, a mere heap, but the whole is something besides the parts [. . .]"*[3].

2.2. Combining Results from Different APIs

We have implemented a wrapper API for the four named entity disambiguation APIs introduced in Subsection 2.1 that returns results in JSON format. While the underlying APIs return entities with their types and/or subtypes, names, relevance, and links into the LOD cloud in different formats, the wrapper API abstracts away the different output formats and returns a common JSON object structure instead. The JSON output for the exemplary micropost *"Tom has the LaTeX, BibTeX, LaTeX, LaTeX blues..."* can be seen in Listing 1. The more APIs agree on a disambiguated named entity, the higher the confidence that (i) the named entity was extracted correctly, and (ii) the named entity was also disambiguated correctly. While in the concrete example "LaTeX" and "blues" were both correctly extracted and disambiguated (`http://dbpedia.org/resource/LaTeX` and `http://dbpedia.org/resource/Blues`), the judgment is based on just one API result in both cases (DBpedia Spotlight and Zemanta respectively), whereas "BibTeX" was correctly extracted and disambiguated (`http://dbpedia.org/resource/BibTeX`) by two APIs at

[3] Aristotle, Metaphysics, Book H 1045a 8-10.



(a) Twitter Swarm NLP

(b) Facebook Swarm NLP

Fig. 3. The two browser extensions in action, displaying the in-page named entity extraction.

the same time (both DBpedia Spotlight and Zemanta). Hence, the confidence is higher in the latter case. The complete named entity reconciliation process is described in [40,48].

```
[
    {
        "name": "LaTeX",
        "uris": [
            {
                "uri": "http://dbpedia.org/
                    resource/LaTeX",
                "source": "spotlight"
            }
        ],
        "source": "spotlight"
    },{
        "name": "BibTeX",
        "uris": [
            {
                "uri": "http://dbpedia.org/
                    resource/BibTeX",
                "source": "zemanta,spotlight"
            }
        ],
        "source": "zemanta,spotlight"
    },{
        "name": "blues",
        "uris": [
            {
                "uri": "http://dbpedia.org/
                    resource/Blues",
                 "source": "zemanta"
            }
        ],
        "source": "zemanta"
    }
]
```

Listing 1: Example JSON output of the named entity disambiguation wrapper, showing different entities and sources.

2.3. *Manipulating Web Pages with Browser Extensions*

Our approach is based on browser extensions. Browser extensions are small software programs written in a combination of HTML, JavaScript, and CSS. For this paper, we focus on extensions based on so-called content scripts. Content scripts are JavaScript programs that run in the context of Web pages via dynamic code insertion. By using the standard Document Object Model (DOM), they can read or modify details of the Web pages a user visits. The advantage of using browser extensions is that the concept is very powerful and generalizable at the same time. Powerful in the sense that it allows for significantly changing one's user experience with social networking sites like Facebook or Twitter and simply adding new features. Generalizable in the sense that the approach is extensible to more social networking sites like MySpace [35], LinkedIn [27], Google+ [18], etc. in the future.

2.4. *Gathering Visitor Data with Web Analytics Software*

In order to gather high-level information on Web page visitors (apart from low-level log file statistics), so-called Web analytics software can be used. Such software is typically implemented by adding an invisible snippet of JavaScript code on the to-be-tracked pages of a website. This code then collects visitor data through requests for a specific 1×1 transparent GIF image, also called Web beacon, that is hosted on a Web

analytics server. During these requests, the page and user data is reported in the query part of the Web beacon's URL. In addition to that, the JavaScript snippet usually sets a first party cookie on a visitor's computer in order to store anonymous information such as the timestamp of the current visit, whether the visitor is a new or returning visitor, and the referrer of the website that the visitor came from. Part of the shared visitor information is the IP address, which allows for IP-based geolocation.

2.5. Pseudocode of the Browser Extensions

In the following Listing 2, we provide the pseudocode of the browser extensions, which helps the reader get a better understanding of the involved flow of data.

```
# initial reporting
report user data to Web analytics tool

# as microposts keep coming in
while true
  for each new micropost on the user's timeline do
    NEs = extract named entities from micropost
    for each named entity in NEs do
      highlight named entity in the micropost
      report named entity to Web analytics tool
    end for
  end for
end while
```

Listing 2: Pseudocode of the browser extensions.

3. Experiment Setup and User Demographics

3.1. Experiment Setup

We initially announced the availability of the extensions via Twitter, Facebook, and on our personal blogs, with the objective of reaching an as broad and unbiased audience as possible. Accumulated click statistics for the announcement links are available via the link shortener service bitly[4]. The extension descriptions contain full disclosure on the collected data and on the usage of a Web analytics tool, however, do not include a concrete mention (apart from a remark on entity ranking), that we use the collected data for an experiment. In addition to that, the extension descriptions do not cross-reference each other, *i.e.*, users are *not* actively encouraged to install both extensions in order to guarantee maximum independence of the experiments.

3.2. User Demographics

As the extensions insert a Web analytics tracking snippet, exact user localization is possible based on the users' current physical location, *i.e.*, completely independent from the origin location users might have registered with Facebook or Twitter, and not to be confused with geotagged microposts. Tables 1a and 1b show the distribution of the top-10 locations of extension users. The complete statistics can be found online[5].

In the period from March 1 to November 8 2011, for the Facebook Swarm NLP, overall *858 unique Facebook users* accessed the extension at least 10 times, in comparison to overall *86 unique Twitter users* for the Twitter Swarm NLP. If we put these figures in contrast to the *seven day active users* statistics for the extensions (Figures 4b and 4a), where the Facebook Swarm NLP reached 135, and the Twitter Swarm NLP 72 *seven day active users* as of November 6 2011, we can derive that overall relatively few Twitter users installed the extension and stayed with it for the whole time of the experiments, whereas overall relatively many Facebook users installed the extension, used it for a short while, and then uninstalled it.

4. Discussion

In this Section, we delve into the collected data, which ranges over more than six months. We cover the period from May 1 to November 12 2011. It is to be noted that our data, while *not* statistically significant, shows promising trends and potential direction for future research. If in the following we present results, those are to be taken with a grain of salt. However, common sense and empiric knowledge suggest that they are correct. The experiments were conducted over more than six months, which reduces the risk of short-term spikes.

[4]Statistics: `https://bitly.com/elP5OW+` (Facebook Swarm NLP) and `https://bitly.com/eBsjQu+` (Twitter Swarm NLP)

[5]Complete statistics: `https://github.com/tomayac/swj-microposts/tree/master/stats`

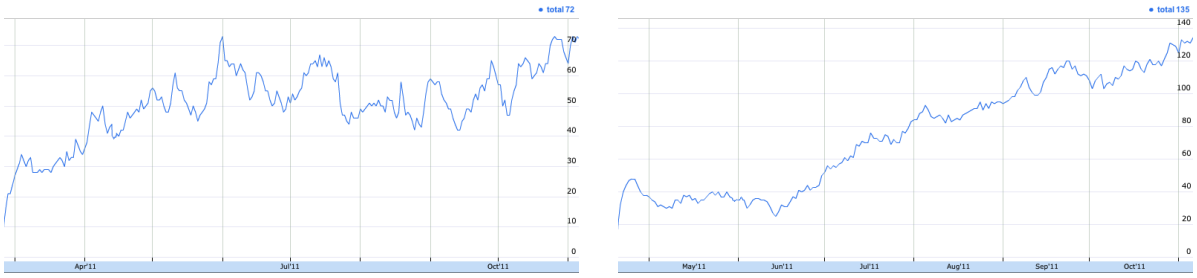
Country	Visits
United States	831
Japan	296
Germany	288
Italy	284
Finland	204
United Kingdom	200
Australia	176
Russia	162
Thailand	155
Spain	147

(a) Twitter Swarm NLP.

Country	Visits
Germany	2,280
Thailand	2,133
Mexico	1,586
Czech Republic	1,416
United Kingdom	1,100
India	1,047
Australia	854
Indonesia	826
Italy	805
United States	742

(b) Facebook Swarm NLP.

Table 1: An analysis of the top-10 locations of the browser extensions' users exposes a varied geographical pattern.



(a) Twitter Swarm NLP

(b) Facebook Swarm NLP

Fig. 4. The seven day active user count for both extensions follows a stable or slowly increasing trend.

4.1. Raw Statistics

First, we present some raw statistics on both social networking sites. Table 2 shows the absolute, total, and unique number of occurrences of named entities from microposts. Interesting here is especially the relation between *unique* named entities and *total* named entities. To clarify the difference between unique and total named entities, we consider the following examples. In a first case, one user during one social networking session reads two different microposts that contain one common named entity (*e.g.*, two consecutive Facebook posts that talk about cats). Here, we would track two total named entities, but only one unique named entity. In a second case, two different users during their social networking session read two different microposts that contain one common named entity. Here, we would still track two total named entities, however, also two unique named entities. In short, a unique named entity is a named entity that during one social networking session of one user appears only once. The number of absolute named entities refers to the absolute distinct number of named entities that ever occurred during the experiments, independent from users and sessions.

4.1.1. Differences in Raw Statistics on Facebook and Twitter

Looking at the numbers, where we have 76.7% unique named entities for Twitter and 43.0% unique named entities for Facebook, we can carefully derive that the reading experience per social networking session on Twitter is more versatile than on Facebook. We need to note, however, that Facebook microposts are generated at a lot lower frequency than Twitter microposts, and that Twitter microposts are limited to 140 characters, which has an impact on both precision and recall of the named entity disambiguation process. The average duration of a visit on Facebook according to our statistics is 66 min. against 28 min. on Twitter, which implies that Twitter users spend less than half the time than Facebook users on their social networking site. Again, our data is at this point *not* statistically significant, especially as we had the interesting phenomenon of almost more than ten times as many abso-

lute distinct Facebook users than Twitter users. However, *seven day active user* statistics show only about double the number of Facebook users (see Section 3).

4.1.2. Statistic Noise Factors

Further research is necessary that takes the following noise factors into account:

- *Micropost length and effect on recall and precision of named entity disambiguation.* Rizzo *et al.* provide first results in this direction [40].
- *Throughput of microposts per time unit (smaller than one day) and effect on versatility of unique named entities per social networking session.* This stands in an interesting contrast to the number of absolute distinct named entities, which is significantly higher on Facebook.
- *Steadiness of the group of experiment participants over time.* The key point of our approach was that participants were unaware that they were contributing to an experiment. We were surprised that so many Facebook users participated, albeit for a short period of time, whereas so few Twitter users participated, however, in the majority during the whole time.
- *Increase reach through focusing on more browser platforms.* Currently we have limited ourselves to one browser platform, however, the approach can be applied to all browser platforms that support extensions or plug-ins.

4.2. Ranking of Named Entities

The core outcome of our experiments is a comparison of the named entities that people *read about* on the social networking sites Facebook and Twitter. From the absolute distinct number of 18,207 (Twitter) and 54,331 (Facebook) of named entities, we have manually cleaned the list of the top-200 named entities on each social network by removing false positives, and normalizing the representing URIs to DBpedia URIs. This was done by inspecting the extracted entities and the words they correspond to in the Analytics tool. In Figure 5, we present the remaining list of top entities for each social network. Visually, the curves are very similar with five named entities occurring many times, and then a long tail of many named entities occurring few times.

Figure 6 zooms in on just the top-10 entities. Interesting to note is the top named entity on each social network, which is the particular social network name "Twitter" and "Facebook" itself.

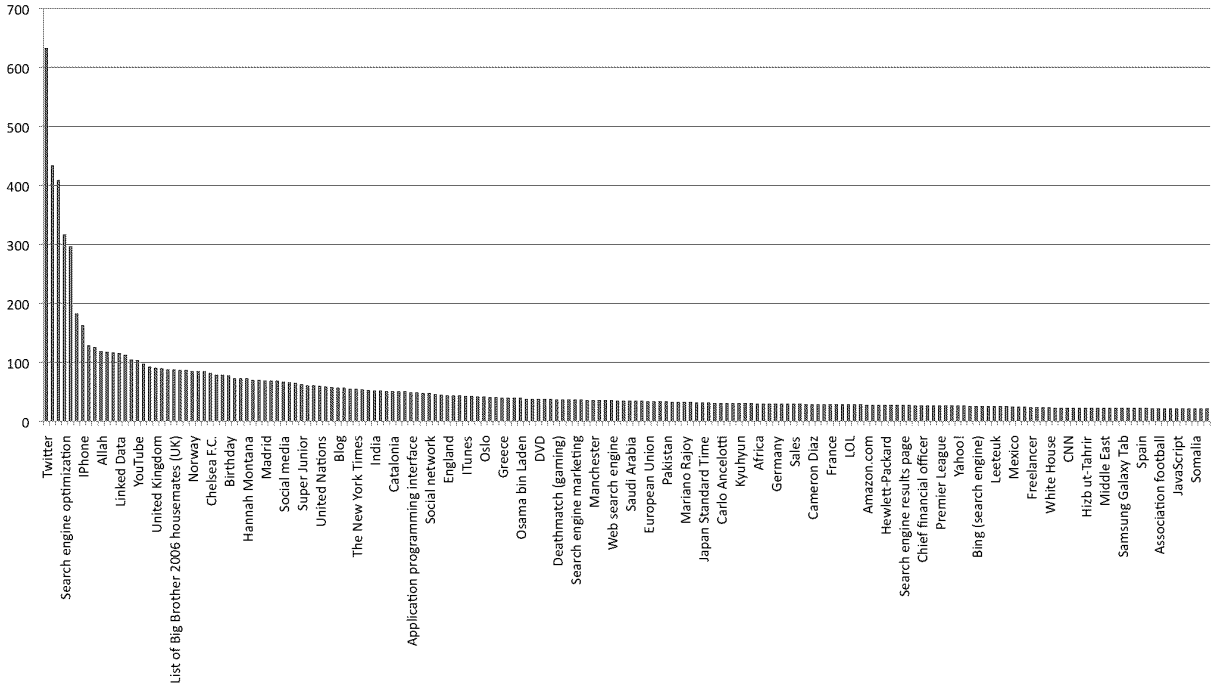
On Twitter, five out of the top-10 named entities are IT-related companies, namely "Twitter", "Google", "Facebook", "Apple", and "Microsoft". Two Apple products, the "iPhone" and the "iPad" hold the positions 7 and 8. "Search engine optimization" (SEO) holds position 4, the "United States" of America are on 5, and position 10 is held by "Allah". Interpreting these results, we can say that 8 out of the top-10 named entities on Twitter by reader interest are of technical nature.

On Facebook, the second most read about named entity is "birthday" on position 2. Position 3 is held by "Allah", who is followed by the (Christian) "God" on position 7. "Love" is on position 4. As the sole company, "Twitter" appears on position 6. The "United States" of America and "North America" are on positions 5 and 9 respectively. The "disc jockey" (DJ) spins a hit single on position 8. Finally, the political organization "United Nations" holds position 10. If we interpret the results, Facebook is used most for personal matters like reading social network friends' birthday felicitations, reading about love and relationships, but also religious matters of Islamic and Christian nature. Music plays a significant role on Facebook, reflected by the presence of disc jockey. With regards to reader interest, there is a tendency towards reading about the United States, or North America in general.

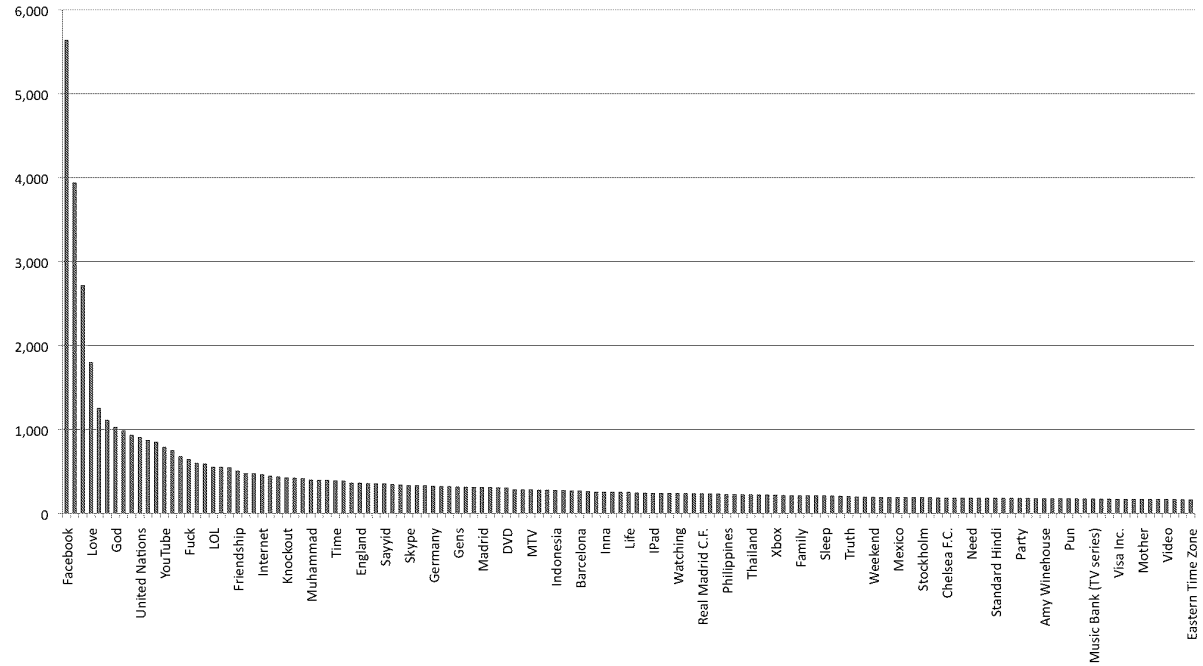
4.3. Segmentation by RDF Type

One of the advantages of using DBpedia URIs to represent named entities in an unambiguous way is that the Linked Data knowledge from DBpedia can be leveraged. Therefore, we have retrieved the RDF type (`rdf:type`) information for the top-500 named entities for both networks, after a manual cleaning operation. Unlike the main DBpedia OWL type (`dbpedia-owl:type`), the `rdf:type` can have multiple values, for example a company can be both a company, and an organization. Type specifications can come from different namespaces, like `dbpedia-owl` (`http://dbpedia.org/ontology/`), `yago` (`http://dbpedia.org/class/yago/`), or `umbel` (`http://umbel.org/umbel/rc/`).

A recent addition is the `schema` namespace `http://schema.org/` around the common schema effort of the big search engines Google, Yahoo!, and Microsoft [19]. The difference lies in the granularity of the underlying ontologies. For example, where `schema` just has "Place", `umbel` differentiates between "Location" and "Populated Place". Applying a



(a) Twitter Swarm NLP.

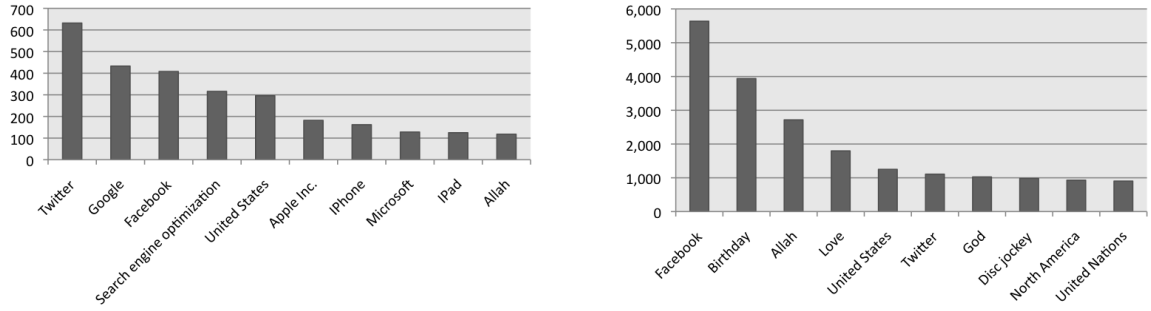


(b) Facebook Swarm NLP.

Fig. 5. The top entities reported by both browser extensions follow a Zipf distribution with a high peak and a long tail.

Network	Absolute	Total	Unique	Unique / Total (%)	Avg. Visit Length
Twitter	18,207	35,958	27,594	76.7%	00:28:12
Facebook	54,331	316,910	136,196	43.0%	01:06:07

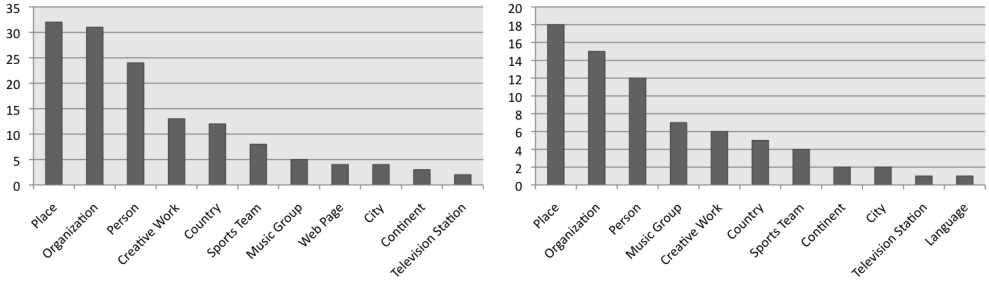
Table 2: Twitter users relatively read more distinct entities than Facebook users do.
However, Facebook sessions take generally longer, which may account for this difference.



(a) Twitter users tend to read technical stories.

(b) Facebook users tend to read personal stories.

Fig. 6. The top-10 consumed entities for both networks clearly show a subject bias.



(a) Twitter

(b) Facebook

Fig. 7. The six-month entity type statistics also follow a Zipf distribution, wherein a few types have many occurrences, and many types have few occurrences. Displayed here are schema.org types, but the same observation can be made for all ontologies and their combinations.

	Total	Northern Africa	Southern Europe	Northern Europe	Northern America	Australia and New Zealand
db:Libya	104	1	38	33	16	9
db:Egypt	86	47	22	0	5	10
db:Syria	69	17	5	8	15	12

(a) Twitter

	Total	Western Europe	Northern Africa	Australia and New Zealand	Northern Europe	South-Eastern Asia
db:Egypt	128	41	47	11	13	3
db:Libya	68	20	20	12	0	5
db:Syria	61	23	9	15	1	5

(b) Facebook

Table 3: Sub continent geolocations of reader interest of some of the Arab Spring key countries (May 1 – November 12).

loose ontology mapping, the top-3 types on both social networks are Place, Organization, Person for `schema` and `dbpedia-owl`, and Place, Organization, Business for `umbel`. Figure 7 shows the distribution of the types of the most common named entities within the observation period, clearly illustrating the long tail of type specifications.

4.4. Events from Recent History

In this Subsection, we present some events from recent history and show how, using our approach, new insights or trends can be revealed. We explicitly highlight the revealed insights by our approach, and compare the limitations of the traditional producer-focused approach.

4.4.1. Comparison of Reader Geolocations for the Arab Spring

The Arab Spring is a revolutionary wave of demonstrations and protests occurring in the Arab world[6]. We compared sub continent geolocations of readers of microposts mentioning the countries of Egypt, Libya, and Syria on both Facebook and Twitter. The results for both social networks can be found in Table 3. First, it is interesting to note the sequence of the countries. On Twitter, it is Libya, Egypt, and Syria, whereas on Facebook it is Egypt, Libya, and Syria. Reader sub continent geolocation distribution is different as well. Where on Facebook the Western European region is most interested in the three countries, on Twitter it is the Northern African region. From Twitter users in Northern Africa, Egypt was the most read about country, in contrast to the among global Twitter users more popular Libya. Note that the use of proxy servers, which was widespread during the event, has at most minimally skewed the statistics, since Analytics takes various factors in account to determine a user's location. Overall, the Arab Spring was almost equally represented on both social networks, albeit our numbers are only sufficient for trend analyses. Both social networks play an important role for the organization of the protests and the distribution of eyewitness statements.

Revealed insight: comparing reader geolocations to distinguish geographic centers of interest can be especially useful to detect local social networking preferences.

Named Entity	Total (in Norway)
db:Oslo	75
db:Norway	57
db:Verdens_Gang	42
db:CNN	19
db:Norwegian_Broadcasting_Corporation	18
db:Jens_Stoltenberg	16

Table 4: The above entities where the most present in user's perception on the 2011 Norway attacks (July 22).

Traditional approach: using the traditional approach of social media mining, it is impossible to limit the analysis to people who physically were in any of the Arab spring countries during that period of time. Alternatives, such as only considering geotagged microposts or using user profile data, are insufficient: consider for instance a foreign journalist who does not geotag microposts and has his profile location set to his hometown. Additionally, with our approach, only microposts that ever appeared on real users' timelines get analyzed, *i.e.*, only microposts that besides being authored also have found an audience.

4.4.2. Facebook Reader Perception of the 2011 Norway Attacks

On July 22, a mass shooting took place on the island of Utøya in Norway, preceded by a car bomb explosion in Oslo[7]. We start our analysis with a deep-dive into the most read about named entities in Norway on the day of the attacks and then manually filter the list for relevance to the event. Table 4 shows the resulting top-6 ranking of named entities on July 22 on Facebook. The list is led by the two geographic entities of the city of "Oslo" and the country of "Norway". Ranks 3 to 5 are held by traditional news media: the popular newspaper "Verdens Gang", typically just referred to as VG, the Cable News Network "CNN", and the "Norwegian Broadcasting Corporation", known as NRK. The current Norwegian Prime Minister "Jens Stoltenberg" follows on rank 6. It is interesting to note how Norwegian readers got detailed information on the attacks through local media (VG and NRK), but also through the international company CNN. Traditional media companies more and more harvest social networking sites for authentic coverage of events. For Utøya, this is documented, *e.g.*, in the case of survivor Adrian Pracon (@AdrianPracon on Twitter) in a tweet

[6] Arab Spring: `http://en.wikipedia.org/wiki/Arab_Spring`

[7] 2011 Norway attacks: `http://en.wikipedia.org/wiki/2011_Norway_attacks`

Shah Alam	Kuching	Kuala Lumpur	Keningau	Kulim
47	34	18	16	16

Table 5: Popularity of the Celcom brand on city level (May 1 – November 12).

from Sky News producer @fimackiesky[8]. While Facebook has terminated terrorist Anders Behring Breivik's profile, the Internet already has conserved a copy[9].

Revealed insight: traditional news media still play the most important role in informing people, albeit the news item itself is shared via social networks.

Traditional approach: analogously to the Arab Spring events, it is impossible to limit the analysis to people who were in Norway at the moment of the attacks. Since the fraction of people that geotag their micropost in Norway is sparse, and using profile location is error-prone, only the reader-focused approach allows to retrieve data from people were in Norway at time of the attacks. We can highlight the posts that affected most Norwegians, *i.e.*, the microposts they have read most.

4.4.3. Brand Popularity on City Level on Facebook

Celcom Axiata Berhad, DBA Celcom, is the oldest mobile telecommunications company in Malaysia[10]. We show the brand's popularity based on named entity occurrences pivoted by cities of readers interested in micropost mentioning the company. Table 5 shows the top-5 cities where people read about Celcom. This allows for targeted brand awareness campaigns in cities where the brand has a low popularity, potentially based on additional sentiment analysis.

Revealed insight: City level analyses allow for fine-grained details on, *e.g.*, brand popularity over time.

Traditional approach: a common occasion for mentioning one's telecommunications provider is to check whether only one's own connection is down, or also everyone else's. Using traditional social media mining, spikes in authoring microposts can be detected. However, the location of consumers potentially affected by such microposts cannot be determined exactly.

[8]Sky News contact Utøya survivor: `https://twitter.com/#!/AdrianPracon/status/94573763500326912`

[9]Copy of the terrorist's Facebook profile: `http://publicintelligence.net/mirror-of-utøya-gunman-anders-behring-breiviks-facebook-page-and-photo-gallery/`

[10]Celcom: `http://en.wikipedia.org/wiki/Celcom`

Named Entity	Total	Unique
db:Cat	161	41
db:Dog	122	43
db:Persian_(cat)	29	5
db:Kitten	23	7
db:Chihuahua_(dog)	9	1

Table 6: Reader popularity of cats and dogs on Facebook (May 1 – November 24).

4.4.4. LOLcats vs. LOLdogs on Facebook

One of the more popular Internet phenomena is the sharing of cute cat and dog photos[11]. On November 23, the well-known link shortening service *bitly* published term popularity-based statistics to test the hypothesis that "kittens really rule the Internet" [8]. According to their results, dogs clearly outperform cats among all *produced* links on microposts. With our approach, we were able to confront bitly's results with the *consumed* microposts around cats and dogs on Facebook. Table 6 shows that among micropost readers, cats indeed haz rule the Internetz.

Revealed insight: the importance of differences between producer and consumer sides becomes evident.

Traditional approach: via traditional social media mining, exclusively the producer side of microposts can be examined. Using our approach, only microposts that ever appeared on real users' timelines get analyzed, limiting the impact of spammers' and/or trend riders' accounts. These accounts typically produce seemingly popular content, which in reality consists of spam messages disguised within trend words, impacting traditional social media mining.

5. Related Work

We report on related work separated into different areas of research:

- *named entities*, which focuses on named entity detection and disambiguation;
- *semantic annotation of microposts*, which focuses on named entity detection and disambiguation specifically in microposts;
- *trend or popularity detection*, which is based on term frequencies;
- *commercialization of social data*, which aims at monetization of gathered insights.

[11]LOLcats: `http://en.wikipedia.org/wiki/LOLcat`

The presented examples are not to be seen as *the* standard selection of relevant work, but rather as representative overview on a plethora of very similar publications and services. In addition to that, we also provide a *comparison of our work to micropost author-focused approaches*.

The Named Entity (NE) recognition and disambiguation task has been addressed in different research communities such as NLP, Web mining and also part of the Semantic Web community. All of them agree on the definition of a Named Entity, which was coined by Grishman *et al.* as an information unit described by the name of a person or an organization, a location, a brand, a product, a numeric expression including time, date, money and percent found in a sentence [20]. One of the first research papers in the NLP field, aiming at automatically identifying named entities in texts, was proposed by Rau [39]. This work relies on heuristics and definition of patterns to recognize company names in texts. The training set is defined by the set of heuristics chosen. Rau's work evolved and was improved later on by Sekine *et al.* [46]. A different approach was introduced when Supervised Learning (SL) techniques were used. The big disruptive change was the use of a large manually labeled datasets. In the SL field, a human being usually trains positive and negative examples to obtain algorithmic classification patterns. SL techniques exploit Hidden Markov Models (HMM) [7], Decision Trees [45], Maximum Entropy Models [9], Support Vector Machines (SVM) [4], and Conditional Random Fields (CRF) [26]. The common goal of these approaches is to recognize relevant keyphrases and to classify them in a fixed taxonomy. The challenges with SL approaches is the unavailability of such labeled resources and the prohibitive cost of creating examples. Semi-Supervised Learning (SSL) and Unsupervised Learning (UL) approaches attempt to solve this problem by either providing a small initial set of labeled data to train and seed the system [22], or by resolving the extraction problem as a clustering one. For instance, a user can try to gather named entities from clustered groups based on the similarity of context. Other unsupervised methods may rely on lexical resources (*e.g.* WordNet), lexical patterns and statistics computed on large annotated corpus [2].

The NER task is strongly dependent on the knowledge base used to train the NE extraction algorithm. Leveraging on the use of DBpedia, Freebase and YAGO, recent methods, coming from Semantic Web community, have been introduced to map entities to relational facts exploiting these fine-grained ontologies. In addition to detect a NE and its type, efforts have been spent to develop methods for disambiguating information unit with a URI. Disambiguation is one of the key challenges in this scenario and its foundation stands on the fact that terms taken in isolation are naturally ambiguous. Hence, a text containing the term `London` may refer to the city `London in UK` or to the city `London in Minnesota, USA`, depending on the surrounding context. Similarly, people, organizations and companies can have multiple names and nicknames. These methods generally try to find in the surrounding text some clues for contextualizing the ambiguous term and refine its intended meaning. Therefore, a NE extraction workflow consists in analyzing some input content for detecting named entities, assigning them a type weighted by a confidence score and by providing a list of URIs for disambiguation. Initially, the Web mining community has harnessed Wikipedia as the linking hub where entities were mapped [24,21]. A natural evolution of this approach, mainly driven by the Semantic Web community, consists in disambiguating named entities with data from the LOD cloud. In [31], the authors proposed an approach to avoid named entity ambiguity using the DBpedia dataset.

Interlinking text resources with the Linked Open Data cloud becomes an important research question and it has been addressed by several services, such as AlchemyAPI, DBpedia Spotlight, Evri, Extractiv, OpenCalais, Yahoo! Term Extraction and Zemanta, which have opened their knowledge to online computation. Although these services expose a comparable output, they have their own strengths and weaknesses but, to the best of our knowledge, few research comparisons have been spent to evaluate them. The creators of the DBpedia Spotlight service have compared their service with a number of other NER extractors (OpenCalais, Zemanta, Ontos Semantic API[12], The Wiki Machine[13], AlchemyAPI and M&W's wikifier [34]) according to an annotation task scenario. The experiment consisted in evaluating 35 paragraphs from 10 news articles in 8 categories selected from the *The New York Times* and has been performed by 4 human raters. The final goal was to create wiki links and to provide a disambiguation benchmark (partially, reused in this work). The experiment showed how DBpedia Spotlight overcomes the performance of other

[12] `http://www.ontos.com`

[13] `http://thewikimachine.fbk.eu/`

services under evaluation, but its performances are strongly affected by the configuration parameters. Authors underlined the importance to perform several setup experiments and to figure out the best configuration set for the specific disambiguation task. Moreover, they did not take into account the precision of the NE and type.

In [41], we proposed a first comparison attempt, highlighting the precision score for each extracted field from 10 news articles coming from 2 different sources, *The New York Times* and *BBC*[14] and 5 different categories: business, health, science, sport, world. Due to the news articles length, we faced a very low Fleiss's kappa agreement score: many output records to evaluate affected the human rater ability to select the correct answer. Indeed, to avoid this problem, Mendes *et al.* proposed a dataset composed of pieces of news articles (paragraphs). Although this approach biases the extraction results for the One Entity per Document extractor, we consider it a valid approximation for the evaluation agreement.

5.1. Semantic Annotation of Microposts

Passant *et al.* introduced a Semantic MicroBlogging (*sic*) framework (SMOB, [38]) that enables a *distributed, open, and semantic* microblogging experience based on Semantic Web and Linked Data technologies by annotating microposts with common vocabularies such as FOAF or SIOC. SMOB relies on distributed autonomous hubs that communicate with each other to exchange microposts and subscriptions, which can also be cross-posted to Twitter. Hashtags, words or phrases preceded by the '#' symbol, have been popularized on Twitter as a way for users to organize and search messages [37]. The authors suggest the use of meaningful hashtags such as `#dbp:Eiffel_Tower` or `#geo:Paris_France`, in the style of widely used RDF prefixes for DBpedia and GeoNames.

In the Linked Open Social Signals project (LOSS, [33]), Mendes *et al.* investigate the representation of microposts as Linked Open Data and address the problem of information overload caused by the sheer amount of microposts (the authors call the opinions, observations, and suggestions contained in microposts "social signals", hence the project name). While the micropost community has come up with hashtags in order to categorize microposts, these hashtags are ambiguous and have to be explicitly added to the micropost by the author and due to length constraints are sometimes left out in favor of more text. The main goal of LOSS is thus to enable collective analysis of social signals for sense-making by using Linked Open Data principles in combination with realtime push models.

5.2. Trend and Popularity Detection

As outlined before, research on trend and popularity detection has mainly focused on Twitter due to the facile availability of data through the Twitter Streaming API. A basic overview is given by by Benhardus in [5], where the author applies and evaluates several methodologies to large Twitter corpora such as (normalized) term frequency, TF-IDF, and entropy. With TwitterMonitor [29], Mathioudakis and Koudas present a *bursty* keyword-based Twitter trend detector demonstration. Their algorithm is able to detect groups of bursty keywords and also enrich trends with potentially associated keywords. Trendsmap [51] provides a realtime mapping of Twitter trends across the world. The service allows for splitting up one's view in different granularity levels by current location, city, region, and world.

What the Trend [58] is a service that provides manually curated and annotated reports on Twitter toptrending topics. The service adds explanations to why tropics trend and data behind trending patterns. Primarily, the information on What the Trend is usergenerated, however, the service also sells curated yearly reports.

Topsy Labs, Inc. offers a commercial API [50] that allows for *applying social intelligence to realtime decisioning*. Therefore, the service applies algorithms that try to rank popular videos, photos, blog posts, and news stories, most influential users on a certain topic, and individual user influence scores. Supported social networks include Twitter and Google+. Different from the social networks themselves, Topsy Labs claims to allow for going back in history up to the year 2008 with their commercial API.

Twimpact [52] is a realtime Twitter data analysis company with special focus on social media communication that reports *immediate events, current trends, and relevant opinion leaders* within social media conversations. Twimpact uses machine learning-based ranking techniques that do not rely on simple reader/-

[14] `http://www.bbc.com`

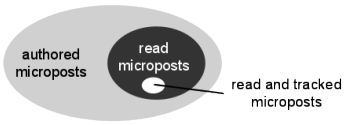


Fig. 8. Only a subset (of unknown size) of all microposts that get authored also ever get read. In turn, we can track only a small subset of all read microposts with our extensions.

follower counts, but rather include actual communication patterns, such as retweets.

5.3. Comparison with Micropost Author-focused Approaches

With our focus on micropost readers rather than on micropost authors, we strive for a paradigm shift with important consequences.

The first and most important consequence is the paradigm shift itself. With our approach, only microposts that ever appeared on real users' timelines get analyzed, *i.e.*, only microposts that besides being authored also have found an audience. Where the traditional social media mining approach in the lack of geotagged microposts has to fall back on user profile data (for example applied in the BreakingNews.com website's submission page `http://www.breakingnews.com/submit#twitter`), with our approach, we can rely on advanced Web analytics data based on very accurate Wi-Fi network and IP-address-based location tracking.

The second consequence is the availability of only a limited set of micropost data, as illustrated by the Venn diagram in Figure 8. Only a subset (of unknown size) of all microposts that get authored also ever gets read, whereof in turn our approach covers an unknown fraction. Where the Twitter Streaming API offers up to 10% coverage of all authored tweets, full access to the "fire hose" with 100% coverage of all authored tweets is handled through Twitter data providers [55]. For Facebook, no such publicly available option exists. For obvious reasons, there is no API from either of the networks for read microposts, which is why we came up with the idea to hook into the reading experience on social networks through browser extensions. Attracting a larger amount of users is feasible if sufficient incentives are provided. For instance, significant extra functionality can be provided on top of the named entity extraction, such as automated micropost summarization to enable a faster and more broad social media experience.

The third consequence has to do with privacy issues. While aggregated public data made available via APIs may feel like violating privacy [47], it is still public data. However, the approach we took in this paper goes one step further by explicitly accessing a social network user's timeline, and reporting back named entities to a Web analytics service (the same that already gets used natively by Twitter). It is important to note, however, that no connection is been made between a user's individual timeline and the person behind. Also, no private conversations are monitored. We have stated all accessed data in the extension descriptions.

The fourth consequence is the potential bias introduced by targeting specific Web browsers with the extensions and the willful neglect of desktop and mobile applications. With regards to specific Web browsers, for now, we have focused on the Google Chrome browser due to its native Chrome Web Store that guarantees optimal exposure of the extensions in a centralized way. Concerning desktop and mobile applications, there is definitely some, albeit unmeasurable, bias. However, unofficial statistics from Twitstat [53] suggest that the desktop Web version of Twitter is still the means for the majority of its users to access the social network. For Facebook, official usage statistics [17] state that from more than 800 million active users more than 350 million currently access Facebook through their mobile devices, which still means that the majority use Facebook via the desktop Web version.

6. Future Work

A drawback of our approach is that getting statistically significant data is difficult. The bigger the so-called panel, the more representative the results. At its core, this is not a new problem. The television viewership audience measurement system described in [25] is comparable to our approach: whenever a viewer switches channels, the system automatically reports a channel switch event back to a tracking server. Analogically, whenever a micropost reader reads a micropost, the detected and disambiguated named entities are automatically reported back to the Web analytics software via our browser extensions. For radio and television audience measurement, a Portable People Meter device [3] developed by Arbitron is used in some parts of the United States. We imagine a similar setting for our approach where randomly selected social network users can be asked to participate in micropost audience measurement studies.

In Subsection 4.1 we have outlined factors to be taken into account in order to improve the quality

of statistical data. With the proposed switch to an officially disclosed micropost audience measurement setup, the two last factors, increasing the study reach and the steadiness of the experiment participants, will resolve nicely in the sense of being the same as with traditional audience measurement.

We have already covered the theoretic consequences of the paradigm shift of focusing on the micropost author side in Subsection 5.3. More work is needed to practically compare the differences in results with common scenarios like the one proposed in the beginning: *"Is my brand X more popular in region R on social network A, or social network B?"*. Evidently, conditions apply, which make the comparison interesting. The determination of the region R has to be interpolated with traditional social media mining, whereas our approach has exact IP geolocation-based location awareness. Reader and author popularity of a brand can be different, and finally, the sample size of the set of examined microposts will be different.

7. Conclusion

Social networks play a crucial role for all sorts of serious and non-serious questions of life. This can go as far as regimes censoring social networks altogether, as it has happened in Egypt [23]. While tech-savvy Internet users can circumvent censorship barriers, the general population is effectively cut off of social network communication. In order to help Egyptians share eyewitness statements about the happenings in their country again, a phone-based *Speak-to-Tweet* service that required no Internet connection was established [49]. Our approach can help prioritize such anti-censorship efforts by analyzing where reader interest is located geographically, and what social networks people use for their information needs.

This paper, to the best of our knowledge, for the first time, was focused on the consumer point of social media. Although the amount of data was smaller than with similar, producer-oriented studies, we were able to confirm some interesting differences in social media consumption. Concretely, the most read Twitter messages are technical in nature, while the most read Facebook updates concern personal matters. This measuring method promises interesting new aspects for future research, for example, to choose the right social medium for a certain data mining task. On the economical side, businesses can use social media consumer behavior to develop advertising strategies, since today, they still depend on producer behavior as a second-degree estimation, or base their decisions on manual and error-prone surveys.

We have presented a generalizable approach towards the comparison of topics people read about on social networks. We have shown how named entity disambiguation combined with classic Web analytics can be applied to the social networks Facebook and Twitter. With concrete examples, we have highlighted how the approach of traditional social media mining can be completed and enriched with our reader-focused approach. We have compared both approaches and worked out the limitations and advantages of both. The main contribution of the paper is on the one hand the comparison of reader topics of two social networks and the classification of those topics, and on the other hand the paradigm shift contained in the approach itself. The approach being generalizable, future studies can cover and compare more social networks.

Acknowledgments

We would like to thank Shaun Roach from AlchemyAPI, Andraž Tori from Zemanta, Pablo Mendes from DBpedia Spotlight, and Tom Tague from OpenCalais for their precious support and/or generous API quota allowances.

T. Steiner is partially supported by the European Commission under Grant No. 248296 FP7 I-SEARCH project. J. Gabarró is partially supported by TIN-2007-66523 (FORMALISM), and SGR 2009-2015 (ALBCOM). The research activities as described in this paper were funded by Ghent University, the Interdisciplinary Institute for Broadband Technology (IBBT), the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research Flanders (FWO Flanders), and the European Union.

References

[1] AlchemyAPI. Available at `http://www.alchemyapi.com/api/entity/`.

[2] E. Alfonseca and S. Manandhar. An Unsupervised Method for General Named Entity Recognition And Automated Concept Discovery. In *1st International Conference on General WordNet*, 2002.

[3] Arbitron. Portable People Meter. Available at `http://www.arbitron.com/portable_people_meters/home.htm`.

[4] M. Asahara and Y. Matsumoto. Japanese Named Entity extraction with redundant morphological analysis. In *International Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL'03)*, pages 8–15, Edmonton, Canada, 2003.

[5] J. Benhardus. Streaming Trend Detection in Twitter. *National Science Foundation REU for Artificial Intelligence, NLP and IR*, 2010.

[6] T. Berners-Lee. Linked Data. W3C Design Issue, July 2006. Available at http://www.w3.org/DesignIssues/LinkedData.html.

[7] D. Bikel, S. Miller, R. Schwartz, and R. Weischedel. Nymble: a high-performance learning name-finder. In *5th International Conference on Applied Natural Language Processing*, pages 194–201, Washington, USA, 1997.

[8] bitly blog. Do kittens really rule the Internet?, Nov. 2011. Available at http://blog.bitly.com/post/13216461842/do-kittens-really-rule-the-internet.

[9] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman. NYU: Description of the MENE Named Entity System as Used in MUC-7. In *7th Message Understanding Conference (MUC-7)*, 1998.

[10] comScore. The Netherlands Lead Global Markets in Twitter.com Reach. comScore Data Mine, Feb. 2011. Available at http://www.comscoredatamine.com/2011/02/the-netherlands-leads-global-markets-in-twitter-reach/.

[11] comScore. Top Global Facebook.com Markets by Percent Reach. comScore Data Mine, Mar. 2011. Available at http://www.comscoredatamine.com/2011/03/top-facebook-markets-by-percent-reach/.

[12] R. Cyganiak and A. Jentzsch. Linking Open Data Cloud diagram. Available at http://lod-cloud.net/.

[13] David Simonds (The Economist). Walled Gardens. Taken from a Presentation "WWW and Hopes for the Future" by Tim Berners-Lee, Feb. 2011. Available at http://www.w3.org/2011/Talks/0222-saudi-tbl/#(25).

[14] N. B. Ellison, C. Steinfield, and C. Lampe. The benefits of Facebook "friends:" social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication*, 12(4):1143–1168, 2007.

[15] Facebook. Available at http://www.facebook.com/.

[16] Facebook. Real-time Updates. Facebook API Documentation, Nov. 2011. Available at https://developers.facebook.com/docs/reference/api/realtime/.

[17] Facebook. Statistics, Nov. 2011. Available at https://www.facebook.com/press/info.php?statistics.

[18] Google+. Available at https://plus.google.com/.

[19] Google, Inc., Yahoo, Inc., and Microsoft Corporation. What is Schema.Org?, 2011. Available at http://schema.org/.

[20] R. Grishman and B. Sundheim. Message Understanding Conference-6: a brief history. In *16th International Conference on Computational linguistics (COLING'96)*, pages 466–471, Copenhagen, Denmark, 1996.

[21] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust Disambiguation of Named Entities in Text. In *Conference on Empirical Methods in Natural Language Processing*, pages 782–792, 2011.

[22] H. Ji and R. Grishman. Data selection in semi-supervised learning for name tagging. In *Workshop on Information Extraction Beyond The Document*, pages 48–55, Sydney, Australia, 2006.

[23] D. Kravets. Twitter blocked in Egypt amid street protests, Jan. 2011. Available at http://edition.cnn.com/2011/TECH/social.media/01/26/twitter.egypt.wired/index.html.

[24] S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti. Collective annotation of Wikipedia entities in Web text. In *15th ACM International Conference on Knowledge Discovery and Data Mining (KDD'09)*, pages 457–466, Paris, France, 2009.

[25] S. Lee, Sanghyeun (Gwangjiu-City, KR), Lee, Byung-tak (Suwon-city, KR). SYSTEM FOR GATHERING TV AUDIENCE RATING IN REAL TIME IN INTERNET PROTOCOL TELEVISION NETWORK AND METHOD THEREOF, Jan. 2010.

[26] A. M. W. Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *7th International Conference on Natural Language Learning at HLT-NAACL (CONLL'03)*, pages 188–191, Edmonton, Canada, 2003.

[27] LinkedIn. Available at http://www.linkedin.com/.

[28] Y. Liu, K. Gummadi, B. Krishnamurthy, and A. Mislove. Analyzing Facebook privacy settings: user expectations vs. reality. In *Proceedings of the 11th ACM/USENIX Internet Measurement Conference (IMC'11)*, Nov. 2011.

[29] M. Mathioudakis and N. Koudas. TwitterMonitor: Trend Detection over the Twitter Stream. In *Proceedings of the 2010 International Conference on Management of Data*, SIGMOD '10, pages 1155–1158, New York, NY, USA, 2010. ACM.

[30] B. Meeder, B. Karrer, A. Sayedi, R. Ravi, C. Borgs, and J. Chayes. We know who you followed last summer: inferring social link creation times in Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 517–526, New York, NY, USA, 2011. ACM.

[31] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia Spotlight: Shedding Light on the Web of Documents. In *7th International Conference on Semantic Systems (I-Semantics)*, 2011.

[32] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia Spotlight: Shedding Light on the Web of Documents. In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)*, 2011.

[33] P. N. Mendes, A. Passant, P. Kapanipathi, and A. P. Sheth. Linked Open Social Signals. *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, 1:224–231, 2010.

[34] D. Milne and I. H. Witten. Learning to link with Wikipedia. In *17th ACM International Conference on Information and Knowledge Management (CIKM'08)*, pages 509–518, Napa Valley, California, USA, 2008.

[35] MySpace. Available at http://www.myspace.com/.

[36] OpenCalais. Available at http://www.opencalais.com/documentation/.

[37] A. Parker. Twitter's secret handshake. The New York Times, June 2011. Available at http://www.nytimes.com/2011/06/12/fashion/hashtags-a-new-way-for-tweets-cultural-studies.html.

[38] A. Passant, T. Hastrup, U. Bojars, and J. Breslin. Microblogging: A Semantic and Distributed Approach. In *Proceedings of the 4th Workshop on Scripting for the Semantic Web, Tenerife, Spain, June 02, 2008, CEUR Workshop Proceedings*, 2008. Available at http://CEUR-WS.org/Vol-368/paper11.pdf.

[39] L. Rau. Extracting company names from text. In *7th IEEE Conference on Artificial Intelligence Applications*, volume i, pages 29–32, 1991.

[40] G. Rizzo and R. Troncy. NERD: Evaluating Named Entity Recognition Tools in the Web of Data. In *ISWC'11, Workshop on Web Scale Knowledge Extraction (WEKEX'11), October 23-27, 2011, Bonn, Germany*, Oct. 2011.

[41] G. Rizzo and R. Troncy. NERD: Evaluating Named Entity Recognition Tools in the Web of Data. In *Workshop on Web Scale Knowledge Extraction (WEKEX'11)*, pages 1–16, Bonn, Germany, 2011.

[42] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 695–704, New York, NY, USA, 2011. ACM.

[43] M. A. Russell. *21 Recipes for Mining Twitter*. O'Reilly Media, 2011.

[44] M. A. Russell. *Mining the Social Web*. Head First Series. O'Reilly Media, 2011.

[45] S. Sekine. NYU: Description of the Japanese NE system used for MET-2. In *7th Message Understanding Conference (MUC-7)*, 1998.

[46] S. Sekine and C. Nobata. Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. In *4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, 2004.

[47] S. Sengupta. When Sites Drag the Unwitting Across the Web. *The New York Times*, Nov. 2011. Available at http://www.nytimes.com/2011/11/14/technology/klouts-automatically-created-profiles-included-minors.html.

[48] T. Steiner, R. Verborgh, J. Gabarró Vallés, and R. Van de Walle. Adding Meaning to Facebook Microposts via a Mash-up API and Tracking Its Data Provenance. In *Proceedings of the 7th International Conference on Next Generation Web Services Practices*, pages 342–345, Oct. 2011.

[49] The Official Google Blog. Some weekend work that will (hopefully) enable more Egyptians to be heard, Jan. 2011. Available at http://googleblog.blogspot.com/2011/01/some-weekend-work-that-will-hopefully.html.

[50] Topsy Labs, Inc. Commercial API, 2011. Available at http://topsylabs.com/products/api/.

[51] Trendsmap. Real-time local Twitter trends, 2011. Available at http://trendsmap.com/.

[52] Twimpact. Real-time social network analysis, 2011. Available at http://twimpact.com/.

[53] Twitstat. Twitter Clients, Aug. 2011. Available at http://www.twitstat.com/twitterclientusers.html.

[54] Twitter. Available at http://www.twitter.com/.

[55] Twitter. Partner Providers of Twitter Data, 2011. Available at https://dev.twitter.com/docs/twitter-data-providers.

[56] Twitter. Streaming API. Twitter API Documentation, Nov. 2011. Available at https://dev.twitter.com/docs/streaming-api.

[57] Venkatesh Rao. The Social Graph as Crude Oil (Go Ahead, Build that YASN!). Forbes Blog, Oct. 2011. Available at http://www.forbes.com/sites/venkateshrao/2011/10/21/the-social-graph-as-crude-oil-go-ahead-build-that-yasn/.

[58] What The Trend. Year In Review, 2011. Available at http://yearinreview.whatthetrend.com/.

[59] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts. Who says what to whom on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 705–714, New York, NY, USA, 2011. ACM.

[60] Zemanta. Available at http://developer.zemanta.com/docs/.