

A Framework for Real-time Semantic Social Media Analysis

Diana Maynard *, Ian Roberts, Mark A. Greenwood, Kalina Bontcheva
Department of Computer Science, The University of Sheffield, Regent Court, 211 Portobello, Sheffield, S1 4DP, United Kingdom
E-mail: {d.maynard,i.roberts,m.a.greenwood,k.bontcheva}@sheffield.ac.uk

Abstract.
This paper presents a framework for collecting and analysing large volume social media content. The real-time analytics framework comprises semantic annotation, Linked Open Data, semantic search, and dynamic result aggregation components. In addition, exploratory search and sense-making are supported through information visualisation interfaces, such as co-occurrence matrices, term clouds, treemaps, and choropleths. There is also an interactive semantic search interface (Prospector), where users can save, refine, and analyse the results of semantic search queries over time. Practical use of the framework is exemplified through two case studies: a general scenario analysing tweets from UK politicians and the public's response to them in the run up to the 2015 UK general election, and an investigation of attitudes towards climate change expressed by these politicians and the public, via their engagement with environmental topics. The paper also presents a brief evaluation and discussion of some of the key text analysis components, which are specifically adapted to the domain and task, and demonstrate scalability and efficiency of our toolkit in the case studies.

Keywords: social media analysis, linked open data, semantic annotation, sentiment analysis, NLP

1. Introduction

Social media is the largest collection of information about society that we have ever had, providing an incredibly rich source of behavioural evidence. However, understanding and using it in a meaningful way is often still a major problem. Gleaning the right information can be tricky because analytics tools either do not provide the right kinds of interpretation, or are simply not accurate, aggregated, enriched or easily interpretable[1]. In the recent 2015 UK elections, for example, numerous analytics tools attempted to understand the attitudes of the public towards the various parties and to predict the outcome of the election, but mostly with quite poor results as they did not take into account many subtle nuances. There are many reasons for this, which are not appropriate to discuss here, but one reason is that it turns out that investigating people's values, and their opinions on specific topics such as the economy, rather than their opinions on particular parties as a whole, gave better insight[2]. Furthermore, simple sentiment analysis tools that look at people's opinions often do not deal well with nuances such as sarcasm, nor the fact that people express their sentiment about very specific events more than about a party overall, which may have subtle differences. We therefore need much more complex forms of analysis in order to understand properly what people are saying.

Social media content is dynamic, reflecting the societal and sentimental fluctuations of the authors. User activities on social networking sites are often triggered by popular or specific events and related entities (e.g. sports events, celebrations, crises, news articles) and

*Corresponding author. E-mail: d.maynard@sheffield.ac.uk.
[1] `http://simplymeasured.com/blog/2015/03/09/5-problems-with-how-marketers-use-social-analytics/`

[2] http://www.theguardian.com/politics/2015/may/14/why-did-the-election-pollsters-get-it-so-wrong

topics (e.g. global warming, terrorism or immigration). One of the main tasks of social media analysis is to gain useful insights into what people are thinking, otherwise known as sentiment analysis or opinion mining. There are many applications of this, such as monitoring political opinions, tracking the influence of public mood on stock market fluctuations, studying the distribution of opinions in relation to demographics, and so on. Understanding what events can cause people to change their opinions, who the leading influencers are, and how opinions change over time are all key areas for current research and, while solutions do exist, are often still far from adequate, with many challenges still to be resolved.

The unique nature of social media data is precisely what makes it also so challenging [8]. It is fast-growing, highly dynamic and high volume, reflecting both the ever-changing language used in today's society, and current societal views. Because Twitter, in particular, is fundamentally a reactive medium (most tweets are responses to recently occurring personal or public events), standard opinion mining tools often do not work well because opinions tend to be event-driven rather than topic-driven. What we mean by this is that people tend not to express generic sentiment on Twitter about topics such as climate change, immigration or upcoming elections, but rather, they express very specific sentiment about a recent or future event (a news headline or newspaper article, a quote from a politician, a job interview, the death of a celebrity, what they had for breakfast, etc.). Best results will thus be obtained for such analytic tools when they are focused on some very specific events and have clear opinion targets. For example, positive responses to a speech expressing a sceptical view of the EU are likely to be demonstrating evidence of negative sentiment towards the EU [49]. Similarly, a tweet "Great post about Scotland!" does not imply any positive sentiment towards Scotland, only towards the post, which might have been positive or negative (or even neutral).

A comparison of social media monitoring tools conducted in October 2014 by Ideya Ltd[3] shows that there are at least 245 tools for social media monitoring available, of which 197 are paid, with the remainder free or using a freemium model. Most of the free tools, at least, do not allow the in-depth and customisable analysis ideally required. Published research has principally concentrated on number-crunching exercises based on topic and entity identification by hashtag, simple keyword or easily available Twitter metadata such as author name, language, number of retweets and so on [8,23,37,34,42]. While some of these methods do involve more complex language processing techniques, these typically comprise simple off-the-shelf sentiment analysis tools such as SentiStrength [47] and SentiWordNet [19] and/or generic basic entity and topic recognition tools such as DBpedia Spotlight [35], or core open source NLP tools such as ANNIE [10], and are not adapted to the domain and task.

As a partial solution to these problems, we present a framework for social media monitoring which combines a series of generic tools inside a flexible architecture that allows each component to be easily adapted to the specific social media monitoring task and its domain. The framework includes data collection, semantic analysis, aggregation, search, and visualisation tools, which allow analysts to dig deep into the data and to perform complex queries which do not just rely on surface information. Furthermore, they enable the analyst to find new and interesting correlations between the data, a task which traditionally can only be done manually and therefore on very small volumes of data, and to view the results in a meaningful way.

The framework is highly scalable and can be used both for off-line processing and live processing of social media. The generic framework and components are described in Section 2. In Sections 3 we show how the toolkit has been adapted to a particular task: the monitoring of political tweets leading up to the UK 2015 elections. This scenario involves both an example of long-term Twitter monitoring and (near)-real time live Twitter stream analysis during a set of televised debates. In Section 4, we provide some examples of queries and findings, respectively. We then describe in Section 5 how the tools have been further adapted to deal with a more sociological analysis of the representation of climate change in politics and of the public's reaction to and engagement with this topic. In Section 6 we present and discuss some evaluation of the analysis tools.

2. An Open Source Framework for Social Media Analysis

The social media analytics toolkit is based around GATE [11], a widely used, open source framework for Natural Language Processing (NLP). The toolkit can perform all the steps in the analytics process: data

[3] `http://ideya.eu.com/reports.html`

collection, semantic annotation, indexing, search and visualisation. In the data collection process, user accounts and hashtags can be followed through the Twitter "statuses/filter" streaming API. This produces a JSON file which is saved for later processing. The tweet stream can also (optionally) be analysed as it comes in, in near real-time, and the results indexed for aggregation, search, and visualisation. Twitter's own "hosebird" client library is used to handle the connection to the API, with auto reconnection and backoff-and-retry.

In the case of **non-live processing**, the collected JSON is processed using the GATE Cloud Paralleliser (GCP) to load the JSON files into GATE documents (one document per tweet), annotate them, and then index them for search and visualisation in the GATE Mímir framework [46]. GCP is a tool designed to support the execution of GATE pipelines over large collections of millions of documents, using a multi-threaded architecture.[4] GCP tasks or batches are defined using an extensible XML syntax, describing the location and format of the input files, the GATE application to be run, and the kinds of outputs required. A number of standard input and output data format handlers are provided (e.g. XML, JSON), but all the various components are pluggable, so custom implementations can be used if the task requires it. GCP keeps track of the progress of each batch in a human- and machine-readable XML format, and is designed so that if a running batch is interrupted for any reason, it can be rerun with the same settings and GCP will automatically continue from where it left off.

In cases where **real-time live stream analysis** is required, the Twitter streaming client is used to feed the incoming tweets into a message queue. A separate semantic annotation process (or processes) then reads messages from the queue, analyses them and pushes the resulting annotations and text into Mímir. If the rate of incoming tweets exceeds the capacity of the processing side, more instances of the message consumer are launched across different machines to scale the capacity.

The live processing system is made up of several distinct components:

– The *collector* component receives tweets from Twitter via their streaming API and forwards them to a reliable messaging queue (JBoss HornetQ). It also saves the raw JSON of the tweets in backup files for later re-processing if necessary.
– The *processor* component consumes tweets from the message queue, processes them with the GATE analysis pipeline and sends the annotated documents to Mímir for indexing.
– Mímir receives the annotated tweets and indexes their text and annotation data, making it available for searching after a short (configurable) delay.

Figure 1 shows the architecture of the live processing system in its simplest form.

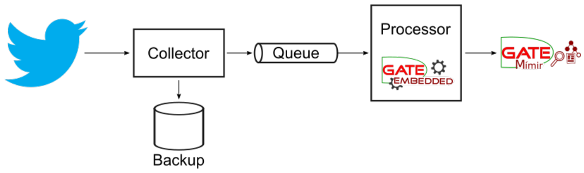

Fig. 1. Simple architecture of live processing system

For the **data collection** component, Twitter offers a set of streaming APIs that deliver tweets to consumers in real time as they are posted. Our system makes use of the statuses/filter API, which allows the user to specify certain constraints and then delivers all tweets (up to a maximum of around 50 per second) that match those constraints. Various kinds of constraints are supported, but the two that are of interest are *track* (a textual filter that delivers all tweets that mention specified keywords, typically hashtags), and *follow* (a user ID filter that delivers all tweets by specified Twitter users, as well as any tweet that is a retweet of, or a reply to, a tweet by one of the specified users). In our political tweets case study described in Section 3, for the live monitoring of debates, we track the hashtags used for each debate, while for the long-term monitoring scenario we simply follow a list of user IDs.

The collector component uses the Hosebird client, a Java library written by Twitter themselves to simplify access to the streaming API. The Hosebird library handles the complexity of authentication, long-lived HTTP connections, and backoff-and-retry behaviour when the connection drops for any reason, so the actual collector logic is very simple. When a tweet arrives on the stream, the collector parses the JSON to extract the tweet ID, then packages the JSON into a message and sends it to the message queue, tagged with its ID (for de-duplication purposes). In parallel, the collector writes the tweet JSON to a backup file, so it is preserved for future reference (for example, if we improve the analysis pipeline we may want to go

[4] For more information about GCP, see `https://gate.ac.uk/gcp/`.

back and re-process previously-collected tweets with the new pipeline). On top of the core collector library, we add a simple web front-end to configure the collector with Twitter API credentials and details of which users and/or hashtags we want to follow.

2.1. Semantic Annotation

GATE has recently been extended to provide numerous tools for social media analysis, namely automatic recognition of terms via TermRaider [12], named entities (people, places, organisations, dates etc.) via TwitIE [7], as well as sentiment analysis (detecting whether a social media post is opinionated, what kind of opinion is expressed, who the holder of the opinion is, what the opinion is about, and so on) [30,29]. Where appropriate, entities and terms are associated with relevant URIs from Linked Open Data via YODIE [22]. TwitIE also comes with a number of general purpose pre-processing components, tailored to social media content, namely Twitter-specific tokeniser, language identifier, normaliser, and POS tagger. Most of these components can (and should) be customised to the domain or application; Section 3 describes how such adaptations have been made for our use case.

The framework also integrates Linked Open Data resources (e.g. DBpedia [5], GeoNames, GEMET), which are accessed via the GraphDB (formerly known as OWLIM) knowledge repository [25]. These are used both during semantic annotation and for semantic search and visualisations, as detailed next.

2.2. Indexing and Querying

Semantic search is more powerful than simple keyword-based search, offering users more precise and relevant results by using the semantics encoded (usually) in ontologies. Google and Facebook refer to such semantics as *knowledge graphs* [45]. Semantic search requires some NLP techniques for understanding word meaning, typically Named Entity Recognition [40] and semantic annotation [6]. The benefit of semantic search, and the grounding of automatically discovered information into ontologies, is that it also enables users to search for knowledge and relationships that are not present in the documents themselves, e.g. which political party an MP represents, so that we can search for all documents written by or which mention MPs from a particular party. It also allows disambiguation of terms: Cambridge, for example, may refer to the city of Cambridge in the UK, to Cambridge in Massachusetts, the University of Cambridge, etc. Similarly, the same concept may be represented by different surface forms, e.g. “the Conservative Party” and “the Tories”.

After analysis, the social media posts are indexed using GATE Mímir [46], which enables complex semantic searches to be performed over the entire dataset. Unlike common search engines such as Google, the query language is not purely keyword based, but instead supports an arbitrary mix of full-text, structural, linguistic and semantic constraints, and can scale to gigabytes of text. Rather than just matching documents in which exact words are to be found, it enables a semantic-based search that can be performed over categories of things, e.g. all Cabinet Ministers or all cities in the UK. Search results can include morphological variants and synonyms of search terms, specific phrases with some unknowns (e.g. an instance of a person and a monetary amount in the same sentence), ranges (e.g. all monetary amounts greater than a million pounds), restrictions to certain date periods, domains etc., and any combination of these. Examples of the kinds of searches that can be performed are given in Section 4.

In terms of the architecture, the processor sends its annotated tweets to a GATE Mímir indexing server. Mímir indexes the plain tweet text, structural metadata like sentence boundaries, hashtags and @mentions, and the semantic annotations detected by the analysis pipeline, such as topic mentions, sentiment expressions, and references to MPs and election candidates. We also index document-level metadata such as the tweet author, the timestamp of the tweet to a suitable level of granularity (the nearest hour for the long-term collection, the nearest minute for the high-intensity debate analysis). In our use case, mentions of candidates and MPs are linked to a semantic knowledge base that provides additional information such as their party affiliation and which constituency they are standing in, and the constituencies are in turn linked to higher-level geographic regions, allowing us to formulate complex queries such as “Find all positive sentiment expressions about the *UK economy* topic in tweets written by Labour candidates for constituencies in Greater London.” By issuing a series of such queries, for each broad topic, party, region and so on, we can generate useful visualizations, as shown in Section 3.

Mímir builds index structures from the annotated data in memory, and performs a “sync to disk” at regular intervals to make the indexed tweets available for processing. The interval between sync jobs determines how close to real-time the tweets become searchable – for the continuous processing of tweets by candidates, one sync per hour is sufficient, but for the debates where we receive thousands of tweets per minute and want to visualise the results as quickly as possible, we sync at least once every five minutes.

2.3. GATE Prospector

The problem of extracting insights from large volumes of social media content is, by its nature, an information discovery task. Such tasks require more sophisticated user interfaces, which enable users first to narrow down the relevant set of documents through an interactive query refinement process, and then to analyse these documents in more detail. These two kinds of actions require corresponding *filtering* and *details-on-demand* information visualisations [44].

Such information discovery and visualisation functionalities are provided by GATE Prospector [46], which is a web-based user interface for searching and visualising correlations in large data sets. Any Mímir indexed data set can be searched with Prospector, and the analyst can easily interrogate the data and identify correlations, providing a visually enhanced understanding of the content. For example, based on the automatically created linguistic annotations, we can discover and visualise the most frequent topics associated with positive or negative sentiment, or which two topics frequently co-occur in a dynamically selected set of documents.

Prospector also supports temporal analytics, such as investigating which topics become more or less popular over a time period, and what events might cause these changes to occur. Prospector can accept exactly the same queries and in the same format as Mímir and shows their results through visualisations. It also has the possibility of enabling canned queries. In Section 4 we will show further examples of data querying and visualisation in Prospector.

2.4. Robustness and scalability

The architecture of the toolkit is deliberately loosely coupled – there is no direct dependency between the collector and processor components, communication being mediated through the message queue – and the components can be distributed across different machines for higher performance and/or robustness. If a processor fails, incoming tweets will simply stack up in the message queue and will be dealt with when the processor restarts.

If the throughput is higher than a single processor can sustain, then one can simply scale out horizontally by starting up more processor instances, and the message queue will handle the sharing out of messages among consumers without duplication. For extremely high throughput, beyond that which a single Mímir instance can handle, each collector could post its annotated tweets to a separate Mímir index, with searches handled through a federated front-end index. However, this has not proved necessary in our tests, as one Mímir instance can easily sustain 10-15,000 tweets per minute, far more than the Twitter streaming API is prepared to deliver.

On the collector side, it is possible to run several collector instances on different machines, all delivering messages to the same queue. These could be clones, all configured to stream the same tweets (to guard against the failure of a single collector), or each collector could be set up to follow a different hashtag (to get around the rate limits Twitter imposes on a single streaming connection). Either way, the message queue takes care of filtering out duplicates so that each distinct tweet is only processed once. This was a factor in the choice of HornetQ as the message broker, as it has native support for duplicate message detection.

2.5. Availability

The core components of the system are open source and freely available as part of GATE via the LGPL licence, and can be downloaded from `http://gate.ac.uk/download`. This includes all the core analysis tools including TwitIE, but some of the domain-specific customisations we made for the political use case are not publicly available. However, the main tools can easily be customised by the user for their own applications as they see fit; the idea behind the toolkit is as a generic framework that can be used for different domains and datasets. Furthermore, some demos and visualisations from the use case are available, and links are given where appropriate throughout the paper.

3. Analysis of Political Tweets

This section describes the application and adaptations of the social media analytics framework to two related real world scenarios: the long-term monitoring of tweets by UK MPs and parliamentary candidates (and responses to those tweets) throughout the 2015 election campaign, and short-term intensive monitoring of tweets with particular hashtags during the televised leaders' debates during the same period. The case study was part of the Political Futures Tracker project, carried out in collaboration with Nesta.[5] A series of blog posts was produced by Nesta during the election period, describing how the toolkit was used to monitor the election, and showing visualisations and discussions of some of the analysis produced.[6]

3.1. Data collection and annotation

We created a corpus by downloading tweets in real-time using Twitter's streaming API, as described in the previous section. The data collection focused on Twitter accounts of MPs, candidates, and official party accounts. We obtained a list of all current MPs[7] and all currently known election candidates[8] (at that time) who had Twitter accounts (506 MPs and 1811 candidates, of which 444 MPs were also candidates). We collected every tweet by each of these users, and every retweet and reply (by anyone) starting from 24 October 2014.

For the purposes of our experiments described in this and the following section, we used a subset of the collection, up until 13 February 2015 (1.8 million tweets, of which approximately 100k are original tweets, 700k are replies, and 1 million are retweets). Candidate-authored tweets were only collected from 13 January onwards, as sufficient information about candidates was unknown prior to this date.

The semantic analysis pipeline consisted of the following components (where not explicitly stated otherwise, these were developed specifically for this political application). **Named Entity Recognition**, using TwitIE [7], identifies Persons, Places, Organisations etc. and **Named Entity Linking**, using YODIE [22], maps these to their respective URIs in Wikipedia or other web-based knowledge sources. Just detecting and classifying these Named Entities is not, however, sufficient, as we also need to detect some specific categories of Person entities in order to understand the opinions of specific people. **MP and Candidate recognition** detects mentions of MPs and election candidates in the tweet - by name or twitter handle - and links them to their respective URIs in DBpedia and YourNextMP. This linking process is explained more fully in Section 3.2. **Author recognition** detects who the author of the tweet is, and links them to the relevant URI as before.

Topic Detection finds mentions in the text of major topics and subtopics, e.g. environment, immigration etc. in various lexical forms, e.g. "fossil fuels" are an indicator of an "environment" topic. The list of topics was derived from the set of topics used to categorise documents on the gov.uk website[9]. Topic detection is performed by means of gazetteer lists for each topic, manually created and then extended semi-automatically. For example, a list for "environment" might contain terms like "climate change", "global warming", "fossil fuels" and so on. Terms are matched in the text under any morphological variant, e.g. singular and plural forms, different verb forms and so on. Since we cannot expect to list all possible ways in which such topics can be expressed, we also match hyponyms, hypernyms and variants of these lists, using rules to associate head terms and modifiers. For example, a hyponym of a base term could be found by adding a preceding adjective. To prevent overgeneration, we use a stop list of words which should not be used to modify existing terms (e.g. colours, numbers, adjectives denoting emotions and so on). We also extended the lists using the TermRaider term extraction tool[10]. This is not used on its own to generate terms, because initial experimentation showed that it overgenerated, even when the cutoff point was set quite high. Instead, we ran it over a large corpus of tweets and extracted the top 250 terms, then manually analysed this list and added any relevant terms to the correct list. Hashtag pre-processing was added, in order to re-tokenise hashtags according to their constituent words [31]. This enables, for example, the term "palm oil" to be matched against the text "#palmoil". This hashtag decomposition is also used in the sentiment analysis component to recognise sentiment-containing hashtags.

Sentiment Analysis detects whether each tweet conveys sentiment and if so, whether it is positive or negative, the strength of this sentiment, and whether

[5] http://www.nesta.org.uk

[6] http://www.nesta.org.uk/blog/introducing-political-futures-tracker

[7] From a list made publicly available by BBC News Labs, which we cleaned and verified, and have now made available at https://gist.github.com/greenwoodma/

[8] List of candidates obtained from https://yournextmp.com

[9] e.g. https://www.gov.uk/government/policies

[10] https://gate.ac.uk/projects/arcomem/TermRaider.html

the statement is sarcastic or not. It also detects who is holding the opinion and what topic the opinion is about, e.g. David Cameron (holder) is being positive (sentiment) about the environment (opinion topic). The sentiment analysis tools were adapted from those developed previously in [29,31], in order to relate specifically to the political tweets scenario. The main adaptation was to capture the fact that we wanted to recognise opinions only when expressed specifically about one of the topics recognised or about another politician or political party. The default sentiment analysis tools recognise opinions about any entity, term or event.

3.2. Linking Open Data

While a number of interesting analyses can be performed over the raw processed data, the scope for discovering interesting connections is greatly widened when the data is made easily searchable. As described in Section 2.2, GATE Mímir is used to index the semantically annotated documents and to allow Linked Open Data to be used to restrict searches. In this use case, the intention was to use DBpedia as a rich source of knowledge that could be used to aggregate information from the individual documents in interesting ways.

For the domain of UK politics, DBpedia contains a wealth of useful information. Every current UK MP is represented, along with their constituency and the political party to which they belong. For geographical information, we make use of the NUTS1 regions. NUTS (Nomenclature of Territorial Units for Statistics) is a geocode standard for referencing the subdivisions of the UK and other EU countries for statistical purposes, and is represented in DBpedia. At the first level (NUTS1), there are 12 UK regions, which we use in order to make geographical observations and visualisations when constituency offers too fine-grained a distinction.

As mentioned in Section 3.1, we have used data from a number of sources to annotate documents, and these same sources were also used to enrich DBpedia with relevant and reliable domain information. The main problem we had to overcome is that there is no single canonical source that covers all existing MPs and candidates for the upcoming election. Instead, we currently have three different sources of data that describe them; DBpedia, Twitter and YourNextMP. All three sources provide URIs that can identify a single person, be that a traditional URI such as provided by DBpedia, or a Twitter handle which can easily be converted to a URI. Each MP and candidate may be described in all three data sources, but will be contained in at least one. Where a person appears in more than one source, we have asserted `owl:sameAs` properties between them in the ontology to ensure that, regardless of which URI is used, all data we have about a person will be available for use at both indexing time and during subsequent semantic searches and aggregation.

Fortunately, each constituency in the UK does have a URI within DBpedia, which we have used as the canonical reference. Information about a constituency contains details of the current MP, but not the candidates known to be standing in the forthcoming election. We have added the information using the `http://nesta.org.uk/property/candidate` property to link URIs for candidates from the YourNextMP dataset to the constituencies within DBpedia.

While aggregation at the level of constituencies is interesting, more useful is to look at the NUTS1 regions. Unfortunately while the regions themselves are present in DBpedia, there is no reliable and consistent way of determining which region a constituency is a member of, so we have again augmented DBpedia to provide this data using the `http://nesta.org.uk/property/partOf` property to model the relationship. Another DBpedia inconsistency is the fact that within the 12 NUTS1 regions there is no way of determing the ID of the region (a three letter code); for some regions this is encoded using the `http://dbpedia.org/property/nutsCode` property, while some use `http://dbpedia.org/property/nuts`, and some do not include the code at all. For consistency we have added the code to all 12 regions using the `http://nesta.org.uk/property/nuts1code` property. The dataset will shortly be made available for public use at `https://gist.github.com/greenwoodma/`.

This data cleaning and linking of sources gives us a rich data set that can be used to restrict search queries in many different ways to produce insightful analysis. For example, Figure 2 shows a query executed in Mímir to find all tweets by Conservative MPs or election candidates that mention something related to the UK economy, and an example of a tweet found. Neither the fact that the tweet author (Richard Short) is a Conservative MP, nor the words UK economy, are explicitly mentioned in the text: the MP information comes from querying DBpedia, while the relationship between "pension" in the text and the economy comes from our semantic annotation. We should note here that the search interface is not particularly user friendly, especially if SPARQL queries are necessary; front-ends can, however, easily be built on top of the generic search interface which are easier for non-expert users. An example of such a front-end for querying news can be seen at `http://demos.gate.ac.uk/pin/`.

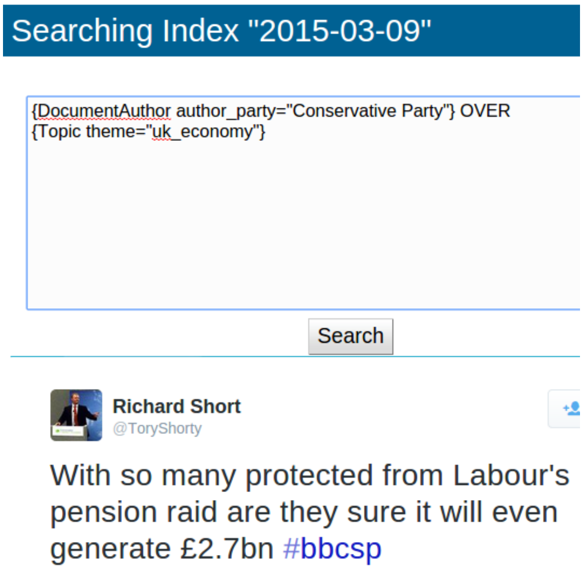

Fig. 2. Example of a Mímir query and result

4. Querying the Data

This section describes how the framework was used to perform a number of search and aggregation queries over the Twitter data, in order to obtain answers to questions such as: how frequently politicians were tweeting, what they were tweeting about, and how this varied between different political parties, between MPs and new election candidates, by region, etc.

A first simple experiment involved aggregating the number of tweets by MPs and candidates by party, based on the DBpedia information of which politician belonged to which party. We found that the Labour Party tweeted more than twice as much as any other party (more than 22,000 tweets, with the next highest being the Conservatives with just over 11,000 tweets). However, when these numbers are normalised by the number of MPs/candidates who had Twitter presence in each party, results showed that Labour MPs had the second lowest proportion of tweets per tweeting MP (average 43.47) with Conservatives lowest at 24.48. In contrast, the smallest parties with the fewest MPs actually had the highest proportion of tweets per tweeting representative: Plaid Cymru, who have only 2 tweeting MPs, had an average of 110 tweets per MP, with the SNP next highest at an average of 85.83 tweets (and 6 tweeting MPs).

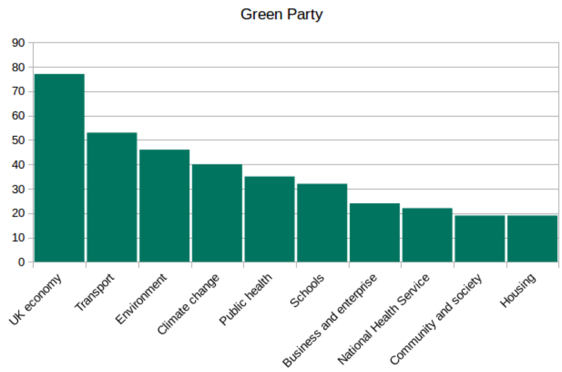

Fig. 3. Top 10 topics mentioned by MPs from the Green Party

We then investigated which topics were mentioned by which party, which uncovered some slightly unexpected results. Figure 3 shows the top 10 topics mentioned by MPs from the Green Party. In order to extract this information, a number of Mímir queries are used, where the party name and topics are varied:

```
{DocumentAuthor author_party =
"Green Party"}| OVER
{Topic theme = "uk_economy"}
```

The information about which party the tweet author belongs to is added automatically from DBpedia during the semantic enrichment phase. The terms are also discovered automatically via the components described in Section 3. The resulting aggregated data is exported in spreadsheet format and charts, and D3-based visualisations are generated from these, e.g. the treemap visualisation shown in Figure 4.

In order to show correlations between parties and topics, we can also use Prospector, which gives us a slightly different way of querying and visualising the results. Figure 5, for example, shows the general purpose UI for exploring associations between semantic annotations/words within a dynamic set of documents returned by a Mímir semantic search query. In this example, two sets of semantic annotations (political topics vs UK political parties in this case) are mapped to the two dimensions of a matrix, while the

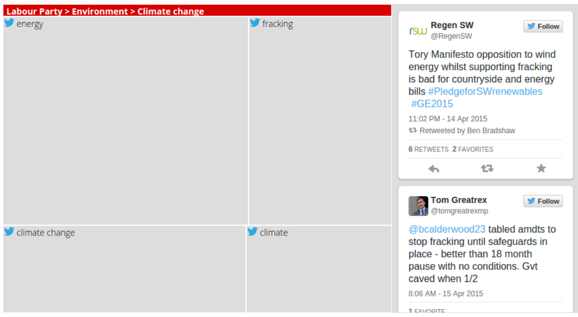

Fig. 4. Treemap showing most frequent terms about climate change mentioned by the Labour Party

colour intensity of each cell conveys co-occurrence strength. The matrix can be re-ordered by clicking on any row/column, which sorts the axis according to the association strength with the clicked item. This example demonstrates the 10 topics most frequently talked about in the run-up to the UK elections in 2015 by the 10 most frequent groups of politicians tweeting, where a group represents a political party and a category (MP or Candidate).[11]

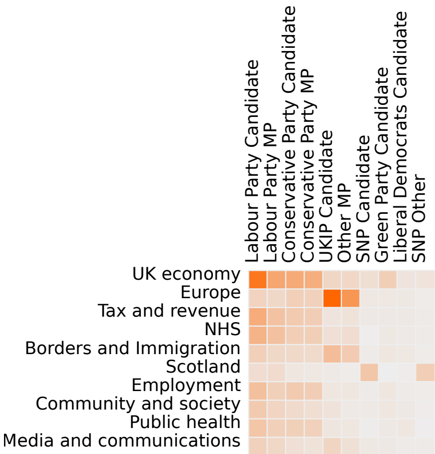

Fig. 5. Prospector's Dynamic Co-occurrence Matrix

Data aggregation can also be carried out on the basis of NUTS regions, not only per party. For instance, it is possible to investigate regional variation of topic mentions, i.e. whether some topics are talked about more in different parts of the country. This involves issuing a series of queries over the tweets for each topic, to find how many tweets mentioning each topic in turn were written by an MP representing each region. The information about which region an MP represents is not expressed in the tweet itself, but uses our knowledge base in two stages: first to find which constituency an MP represents, and then to match the constituency with the appropriate NUTS region, as described in Section 3.2. Figure 6 shows a choropleth depicting the distribution of MPs' tweets which discuss the UK economy (the most frequent theme) during the week beginning the 2nd of March 2015. This is a dynamic visualisation, based on the Leaflet library[12] and the aggregated query results returned by Mímir for each theme and NUTS1 region. The choropleth has a pull-down menu from which the user can select the topic of interest, and this re-draws the map accordingly. Demos of the interactive choropleth and treemap on this dataset, as well as examples of the topic cloud and a sentiment visualisation, are publicly available at `http://www.nesta.org.uk/blog/4-visualisations-uk-general-election`.

It is also possible to query and visualise a dynamically changing subset of matching tweets in Prospector, to uncover patterns in the data. Figure 7, for example, shows the top 20 topics mentioned by MPs and candidates from the Sheffield Hallam constituency. This is the result of a semantic search via a SPARQL query in Mímir, which returns all tweets authored by MPs and candidates from that constituency. On this dynamically selected tweet subset, Prospector then builds frequency and co-occurrence statistics for the selected semantic annotation type (topics in this case). In our example, the most frequently mentioned topics are displayed both as a list and as a term cloud. Note that because Prospector is rather complicated and requires some training to use, it is not currently available publicly as a demo.

5. Measuring Climate Change Engagement

In our other (related) use case, we wanted to investigate how people engage specifically with climate change in politics. Scientists predict adverse consequences unless stronger actions against climate change are taken, but collective awareness about many climate change issues is still problematic. One reason is that people are exposed to vast amounts of conflicting in-

[11] "SNP Other" denotes the odd case where the leader of the SNP party was not an MP or candidate, but was still interesting enough for us to follow. "Other MP" denotes MPs from the minor political parties.

[12] `http://leafletjs.com/`

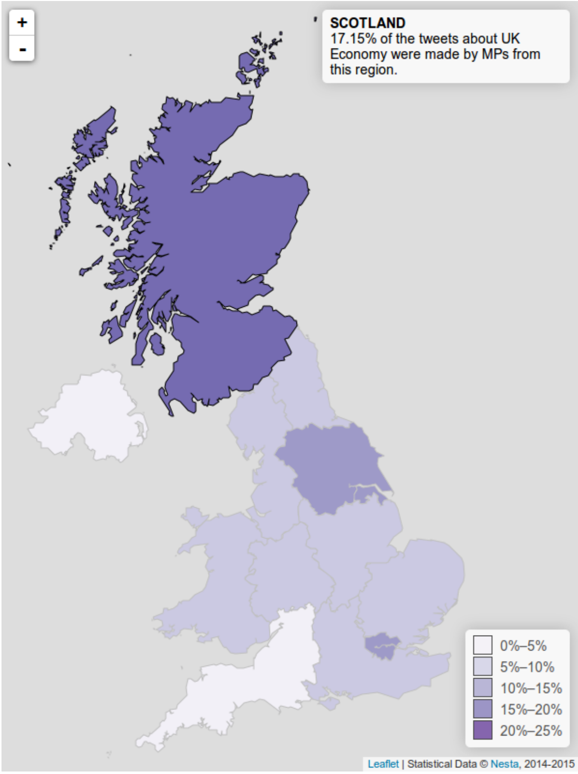

Fig. 6. Choropleth depicting distribution of tweets about the economy

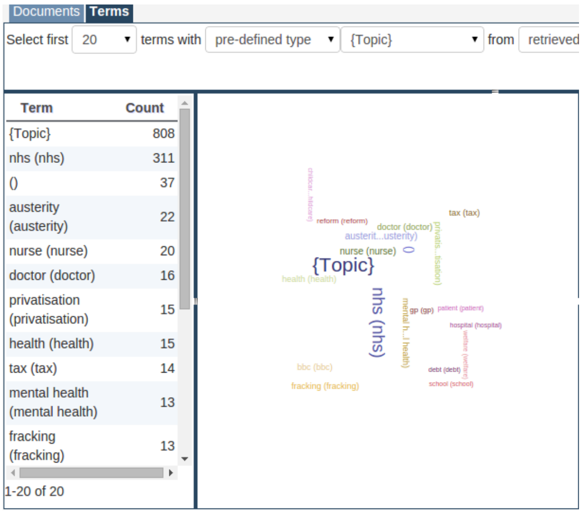

Fig. 7. Top terms mentioned by MPs and candidates from the Sheffield Hallam constituency

formation, making it hard to know what is accurate and relevant. The EU DecarboNet project[13] aims to help solve this problem by developing tailored information services to help empower citizens. Recent studies indicate that a growing awareness about climate change not only results in changes in individual consumption behaviour, but also in individuals engaging more with politics in order to instigate the changes they believe are necessary. In a world where political disengagement is pervasive, this presents an interesting phenomenon. We performed an analysis of political tweets about the environment as part of some work exploring why climate change is seemingly resulting in engaged citizens, when so many other issues seem to leave the public cold and apathetic [16]. We therefore used our political tweets dataset described above in order to try to understand engagement of the public with respect to the topic of climate change and the environment.

We measured engagement with the different political topics described in Section 3 in four ways. First, we looked at retweets. On Twitter there are two main ways in which engagement with tweets is typically measured: retweets and favourites [34]. The problem with the favourite button[14] is that it is used for a variety of different purposes, not just as a means of showing support [33]. In their study, only 65% of study participants knew it existed, and of these only 73.5% ever used it. Retweeting is generally a better way to measure engagement, for a number of reasons. As with favouriting, people retweet for a number of reasons, including both self-gain and philanthropy[9]. By its nature, retweeting constitutes a stronger form of engagement in that it disseminates a tweet to a wider audience and thus propagates a message faster and more globally. Favouriting, on the other hand, is a more private form of approval, since it is a form of interpersonal rather than mass non-verbal communication. Also, on a practical level, if one is collecting tweets via the Twitter streaming API, then one cannot collect favourites as this information is unknown at the time of tweeting. Retweets, on the other hand, can be collected as they count as individual tweets, whereas favourite information is only registered as a count on the original tweet and does not constitute a new tweet. Retweeting is also considered a social action - people typically think explicitly about their followers when tweeting and retweeting. Even though users more than one step away will be unknown to them, they usually have some idea in mind of what kind of people will be in these networks and what their interests will be. Retweeting

[13] http://www.decarbonet.eu

[14] now rebranded as a "Like" button

can thus be seen as a kind of crowdsourcing mechanism. However, this works best when the author is highly influential, for example, a politician, pop star, or other famous person.

We found a high number of climate change related retweets, which indicates a high level of engagement according to the criteria discussed above. 64.48% of the climate change tweets in our dataset were retweets, and 94.3% of them were either retweets or replies. On the other hand, the percentage was much higher than for many other topics such as schools (57% retweets, and 90% retweets and replies).

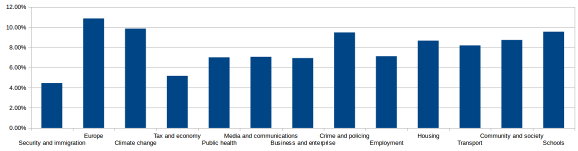

Fig. 8. Percentage of opinion-bearing tweets per topic

Second, we looked at sentiment, which has previously been shown to be a good indicator of engagement [34]. Figure 8 illustrates the percentage of opinionated tweets for each topic. Here we see that climate change is the second highest, after only Europe. We also investigated what percentage of retweets were opinionated (3rd highest), what percentage of opinionated tweets were retweeted (5th highest), what percentage of opinionated tweets were retweets or replies (3rd highest), what percentage of optimistic tweets were retweeted (4th highest, with "Employment" being top) and what percentage of opinionated retweets were optimistic as opposed to pessimistic (2nd highest after "Schools"). This high level of sentiment-filled tweets and retweets about climate change in comparison to other political issues is an indication of a high level of engagement.

Third, we looked at how many tweets contained a mention of another user, since this has also proven to be a good indicator of engagement [34]. Again, climate change scored 3rd highest (after "business and enterprise" and "schools"). Finally, we investigated the number of URLs found in climate change tweets. In Boyd's study of random tweets [9], 52% of retweets contained a URL. This is important because it tells us something about the nature of tweets that engage people (i.e. original tweets containing a URL are more likely to be retweeted). In our corpus, tweets about climate change had the highest percentage of URLs (62%) with the next highest being the topic of schools (56%). Interestingly, 51.4% of climate change retweets contained a URL, while only 45% of retweets about schools contained one. This reveals something about the nature of the engagement: if individuals retweet or reply to such posts, it can be assumed that most of these individuals will further engage by following the link and reading material around the subject of climate change.

Our analysis revealed that climate change and related topics, while not mentioned frequently by politicians other than by the Green Party and UKIP candidates, have a high level of engagement by the public. Although climate change still has a slightly lower engagement rate than topics such as Europe and the economy, engagement still ranks very highly, mostly residing in the top three of most engaged topics.

6. Evaluation

While the analysis toolkit has many interesting features and can provide valuable insights into social media (and other) data, the results are of course only meaningful if the analysis tools perform well. The NLP processing components are thus critical: if entities, topics and sentiments are not extracted correctly, the results are at best meaningless and at worst, could even be highly misleading. One must always bear in mind, however, that tools for automatic text analysis are never perfect, as language is highly ambiguous even in well-written texts such as news reports, let alone noisy text such as tweets [27,14]. However, in large-scale analyses, a few individual errors are not generally problematic as long as the overall trend is correct – for example, if one is analysing changes in sentiment over time with respect to a particular topic or person, as long as the majority of tweets are correctly annotated then the trend will be the same.

The various linguistic analysis tools have been evaluated individually, at least for the core components if not specifically for the adapted versions. The Named Entity Recognition component TwitIE has been evaluated favourably in [7], and performed better than two state-of-the-art Twitter-specific systems, Stanford-twitter [20] and a tool developed by Ritter [41], achieving 80% F1-measure on a corpus of tweets. The Named Entity Linking and Disambiguation component YODIE has been evaluated in [13] against two state-of-the-art tools DBpedia Spotlight [35] and Zemanta[15], achieving the highest Precision (67.59%) and

[15] `http://www.zemanta.com`

F1 score (45.20%). While this is admittedly not that high, these figures are much improved when operating in a narrow domain such as our political tweets set, as ambiguity is considerably reduced, improving Precision, as are the kinds of entities we are interested in, which improves Recall.

An earlier version of the environmental term recognition component has been evaluated in [32] and showed promising results. On a corpus of climate change tweets, it achieved Precision of 85.87%, Recall of 53.05% and F1 of 65.58%. We expect the results on the political dataset to be higher because since that evaluation we have improved the Recall considerably by adding the term expansion techniques. On that corpus, TwitIE scored a Precision of 85.87%, but again, we would expect the results to be much higher on our political dataset, for the reasons given above. Finally the sentiment analysis has been recently evaluated in [28]. On a corpus of environmental tweets, it achieved accuracy of 86.80%, beating three other state-of-the-art systems DIVINE [21], ARCOMEM [27] and SentiStrength [47]. We would expect performance on the political dataset to be similar; in particular, our sentiment analysis tool covers many issues that others do not, such as more fine-grained analysis, specifically dealing with problems such as sarcasm, and detection of opinion targets and holders. Furthermore, we have shown how it can be adapted to deal with slightly differing tasks, such as explicitly recognising only opinions about certain topics or by certain groups of people.

7. Related Work

The main challenge in analysis and visualisation of high-volume social media content is in providing suitably aggregated, high-level overviews. Timestamp-based list interfaces that show the entire, continuously updating stream (e.g. the Twitter timeline-based web interface) are often impractical, especially for analysing high-volume, bursty events. For instance, during the royal wedding in 2011, tweets during the event exceeded 1 million. Similarly, monitoring long running events, such as presidential election campaigns, across different media and geographical locations is equally complex.

One of the simplest and most widely used visualisations is word clouds. These generally use single word terms, which can be somewhat difficult to interpret without extra context. Word clouds have been used to assist users in browsing social media streams, including blog content [3] and tweets [43,36]. For instance, Phelan *et al* [38] use word clouds to present the results of a Twitter based recommendation system. The Eddi system [4] uses topic clouds, showing higher-level themes in the user's tweet stream. These are combined with topic lists, which show who tweeted on which topic, as well as a set of interesting tweets for the highest ranked topics. The Twitris system derives even more detailed, contextualised phrases, by using 3-grams, instead of uni-grams [36]. More recently, the concept has been extended towards image clouds [17].

The main drawback of cloud-based visualisations is their static nature. Therefore, they are often combined with timelines showing keyword/topic frequencies over time [1,4,24,48], as well as methods for discovery of unusual popularity bursts [3]. [15] use a timeline which is synchronised with a transcript of a political broadcast, allowing navigation to key points in a video of the event, and displaying tweets from that time period. Overall sentiment is shown on a timeline at each point in the video, using simple colour segments. Similarly, TwitInfo [26] uses a timeline to display tweet activity during a real-world event (e.g. a football game), coupled with some example tweets, colour-coded for sentiment. Some of these visualisations are dynamic, i.e. update as new content comes in (e.g. topic streams [17], falling keyword bars [24] and dynamic information landscapes [24]).

In addition, some systems try to capture the semantic relatedness between topics in the media streams. For instance, BlogScope [3] calculates keyword correlations, by approximating mutual information for a pair of keywords using a random sample of documents. Another example is the information landscape visualisation, which conveys topic similarity through spatial proximity [24]. Topic-document relationships can be shown also through force-directed, graph-based visualisations [18]. Lastly, Archambault *et al* [2] propose multi-level tag clouds, in order to capture hierarchical relations.

Opinions and sentiment also feature frequently in social media analytics. For instance, Media Watch [24]) combines word clouds with aggregated sentiment polarity, where each word is coloured in a shade of red (predominantly negative sentiment), green (predominantly positive), or black (neutral/no sentiment). Search results snippets and faceted browsing terms are also sentiment coloured. Others have combined sentiment-based colour coding with event timelines [1], lists of tweets [26], and mood maps [1]. Aggregated sentiment is typically presented using pie charts [48] and, in the case of TwitInfo, the overall statistics are normalised for recall [26]).

Most of these social media search and visualisation methods tend to use shallow textual and frequency-based information. The contribution of our work lies in taking into account the extra semantic knowledge about the entities, terms, and sentiment mentioned in the media streams, based on information from Linked Open Data resources such as DBpedia. This semantic knowledge also underpins the data aggregation (e.g. location-based, party-based) and visualisation UIs. In addition, our framework enables the exploration of media streams through topic-, entity-, and time-based visualisations, which make heavy use of the semantic knowledge. In this respect, our work is similar to the KIM semantic platform, which is, however, aimed at static document collections [39].

8. Conclusions

This paper presented an overview of the GATE-based open source framework for (real-time) analytics of social media, including semantic annotation, search and visualisation components. The framework is independent of the particular application domain, although domain-specific customisations can easily be incorporated through additional content analytics components. Knowledge from Linked Open Data is used to power the semantic searches, as well as as the basis for result aggregation and visualisation. For the latter, we employ both our own information discovery environment (Prospector), as well as web-based visualisations (e.g. choropleths, treemaps), which are generated using the D3 and Leaflet JavaScript libraries.

In order to demonstrate the abilities of the framework, a real-life, political science application was discussed. We looked both at a general analysis of the political discourse in the run up to the 2015 UK general elections, and also at the specific question of understandng the role of climate change in today's political debates. While we were not seeking in this study to predict the outcome of the vote, it turns out in retrospect that the kinds of questions we were able to answer with our analysis did actually point to the correct winners, because we were able to use the tools to focus on things like values and topics that people cared about (both from the public and the politicians' point of view), and focus on region-specific criteria (for example, which topics were most talked about / engaged with in which part of the country, rather than just overall sentiment about which party people felt positive or negative about. As part of the ForgetIT project, this example scenario is currently being extended to cover the House of Commons debates, which will include more information about the political roles MPs fulfil. The aim of this is to investigate the evolution of context in an organizational setting, looking at indicators such as changes to ontologies over time.

In our climate change study, the use of semantic annotation and Mímir allows us to search for environmental terms expressed in a multitude of different ways (thanks to the results from the linguistic analysis), including synonyms and hypernyms of the terms mentioned. Even a non-exert user can easily search for not just a particular politician saying something about climate change, but any Labour MP, based on knowledge about UK MPs, which is encoded formally in DBpedia. Furthermore, the analysis is not limited to searching for relevant documents that match a query, but we can also find answers to questions like "Which political party talks the most about environmental topics?", "Which politician gets the most retweets when he/she talks about climate change?", or "In which area of the country are people most engaged in climate change topics on social media?". These kinds of questions can lead to many further interesting kinds of studies by social scientists, environmentalists and politicians, to name but a few. It is easy to see how such techniques can also be applied to other domains and datasets.

9. Acknowledgments

This work was partially supported by the European Union under grant agreements No. 610829 DecarboNet and 600826 ForgetIT, and by the Nesta-funded Political Futures Tracker project.

References

[1] B. Adams, D. Phung, and S. Venkatesh. Eventscapes: Visualizing events over time with emotive facets. In *Proceedings of the 19th ACM International Conference on Multimedia*, pages 1477–1480, 2011.

[2] D. Archambault, D. Greene, P. Cunningham, and N. J. Hurley. ThemeCrowds: Multiresolution summaries of Twitter usage. In *Workshop on Search and Mining User-Generated Contents (SMUC)*, pages 77–84, 2011.

[3] N. Bansal and N. Koudas. Blogscope: Spatio-temporal analysis of the blogosphere. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 1269–1270, 2007.

[4] M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. H. Chi. EDDI: Interactive topic-based browsing of social status streams. In *Proceedings of the 23nd ACM Symposium on User Interface Software and Technology (UIST)*, pages 303–312, 2010.

[5] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia – a crystallization point for the web of data. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 7:154–165, 2009.

[6] K. Bontcheva and H. Cunningham. Semantic annotations and retrieval: Manual, semiautomatic, and automatic generation. In J. Domingue, D. Fensel, and J. Hendler, editors, *Handbook of Semantic Web Technologies*, pages 77–116. Springer, 2011.

[7] K. Bontcheva, L. Derczynski, A. Funk, M. A. Greenwood, D. Maynard, and N. Aswani. TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. Association for Computational Linguistics, 2013.

[8] K. Bontcheva and D. Rout. Making sense of social media through semantics: A survey. *Semantic Web - Interoperability, Usability, Applicability*, 2013.

[9] D. Boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on System Sciences*, pages 1–10. IEEE, 2010.

[10] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: an Architecture for Development of Robust HLT Applications. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 7–12 July 2002*, ACL '02, pages 168–175, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[11] H. Cunningham, V. Tablan, A. Roberts, and K. Bontcheva. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS Computational Biology*, 9(2):e1002854, 02 2013.

[12] E. Demidova, D. Maynard, N. Tahmasebi, Y. Stavrakas, V. Plachouras, J. Hare, D. Dupplaw, and A. Funk. Extraction and Enrichment. Deliverable D3.3, ARCOMEM, 2013.

[13] L. Derczynski, D. Maynard, N. Aswani, and K. Bontcheva. Microblog-Genre Noise and Impact on Semantic Annotation Accuracy. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*. ACM, 2013.

[14] L. Derczynski, D. Maynard, G. Rizzo, M. van Erp, G. Gorrell, R. Troncy, and K. Bontcheva. Analysis of named entity recognition and linking for tweets. *Information Processing and Management*, 51:32–49, 2015.

[15] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*, pages 115–122, 2010.

[16] A. Dietzel and D. Maynard. Climate change: A chance for political re-engagement? In *Proc. of the Political Studies Association 65th Annual International Conference*, 2015.

[17] M. Dork, D. Gruen, C. Williamson, and S. Carpendale. A visual backchannel for large-scale events. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1129–1138, Nov. 2010.

[18] J. Eisenstein, D. H. P. Chau, A. Kittur, and E. Xing. Topicviz: Semantic navigation of document collections. In *CHI Work-in-Progress Paper (Supplemental Proceedings)*, 2012.

[19] A. Esuli and F. Sebastiani. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC 2006*, 2006.

[20] J. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics, 2005.

[21] S. Gindl, A. Weichselbraun, and A. Scharl. Cross-domain contextualisation of sentiment lexicons. In *Proceedings of 19th European Conference on Artificial Intelligence (ECAI-2010)*, pages 771–776, 2010.

[22] G. Gorrell, J. Petrak, K. Bontcheva, G. Emerson, and T. Declerck. Multilingual resources and evaluation of knowledge modelling - v2. Technical Report D2.3.2, Trendminer Project Deliverable, 2014.

[23] K. Holmberg and I. Hellsten. Analyzing the climate change debate on twitter âĂŞ content and differences between genders. In *Proceedings of the ACM WebScience conference*, pages 287–288, Bloomington, IN, USA, 2014.

[24] A. Hubmann-Haidvogel, A. M. P. Brasoveanu, A. Scharl, M. Sabou, and S. Gindl. Visualizing contextual and dynamic features of micropost streams. In *Proceedings of the #MSM2012 Workshop, CEUR*, volume 838, 2012.

[25] A. Kiryakov. OWLIM: balancing between scalable repository and light-weight reasoner. In *Proceedings of the 15th International World Wide Web Conference (WWW2006), 23–26 May 2010*, Edinburgh, Scotland, 2006.

[26] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. TwitInfo: Aggregating and visualizing microblogs for event exploration. In *Proceedings of the 2011 Conference on Human Factors in Computing Systems (CHI)*, pages 227–236, 2011.

[27] D. Maynard. Challenges in Analysing Social Media. In G. S. Adrian DuÅ§a, Dietrich Nelle and G. G. Wagner, editors, *Facing the Future: European Research Infrastructures for the Humanities and Social Sciences*. SCIVERO Verlag, Berlin, 2014.

[28] D. Maynard. Environmental opinion extraction. Technical Report D2.3.1, DecarboNet Project Deliverable, 2015.

[29] D. Maynard, K. Bontcheva, and D. Rout. Challenges in developing opinion mining tools for social media. In *Proceedings of @NLP can u tag #usergeneratedcontent?! Workshop at LREC 2012*, Turkey, 2012.

[30] D. Maynard, G. Gossen, M. Fisichella, and A. Funk. Should I care about your opinion? Detection of opinion interestingness and dynamics in social media. *Journal of Future Internet*, 2015.

[31] D. Maynard and M. A. Greenwood. Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. In *Proceedings of LREC 2014*, Reykjavik, Iceland, 2014.

[32] D. Maynard and A. Scharl. Text analytics tools for environmental information extraction. Technical Report D2.2.1, DecarboNet Project Deliverable, 2014.

[33] F. Meier, D. Elsweiler, and M. L. Wilson. More than liking and bookmarking? Towards understanding twitter favouriting behaviour. In *Proceedings of the International AAAI Conference*

on Weblogs and Social Media, 2014.

[34] C. Meili, R. Hess, M. Fernandez, and G. Burel. Earth hour report. Technical Report D6.2.1, DecarboNet Project Deliverable, 2014.

[35] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia Spotlight: Shedding light on the web of documents. In *Proc. of I-SEMANTICS*, pages 1–8, 2011.

[36] M. Nagarajan, K. Gomadam, A. Sheth, A. Ranabahu, R. Mutharaju, and A. Jadhav. Spatio-temporal-thematic analysis of citizen sensor data: Challenges and experiences. In *Web Information Systems Engineering*, pages 539–553, 2009.

[37] R. Pfitzner, A. Garas, and F. Schweitzer. Emotional divergence influences information spreading in twitter. *ICWSM*, 12:2–5, 2012.

[38] O. Phelan, K. McCarthy, and B. Smyth. Using Twitter to recommend real-time topical news. In *Proceedings of the 2009 ACM Conference on Recommender Systems*, pages 385–388, 2009.

[39] B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, and M. Goranov. KIM – Semantic Annotation Platform. In *2nd International Semantic Web Conference (ISWC2003)*, pages 484–499, Berlin, 2003. Springer.

[40] L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics, 2009.

[41] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: An experimental study. In *Proc. of Empirical Methods for Natural Language Processing (EMNLP)*, Edinburgh, UK, 2011.

[42] M. Rowe and H. Alani. Mining and comparing engagement dynamics across multiple social media platforms. In *Proceedings of the 2014 ACM conference on Web science*, pages 229–238. ACM, 2014.

[43] D. Shamma, L. Kennedy, and E. Churchill. Tweetgeist: Can the Twitter timeline reveal the structure of broadcast events? In *Proceedings of CSCW 2010*, 2010.

[44] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages*, pages 336–343, 1996.

[45] A. Singhal. Introducing the knowledge graph: things, not strings. http://googleblog.blogspot.it/2012/05/introducing-knowledge-graph-things-not.html, May 2012.

[46] V. Tablan, K. Bontcheva, I. Roberts, and H. Cunningham. Mímir: an open-source semantic search framework for interactive information seeking and discovery. *Journal of Web Semantics*, 2014.

[47] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

[48] J. Y. Weng, C. L. Yang, B. N. Chen, Y. K. Wang, and S. D. Lin. IMASS: An Intelligent Microblog Analysis and Summarization System. In *Proceedings of the ACL-HLT 2011 System Demonstrations*, pages 133–138, Portland, Oregon, 2011.

[49] S. Wibberley, J. Reffin, and D. Weir. Method51 for mining insight from social media datasets. *COLING 2014*, page 115, 2014.