
Article:
Maynard, D. G., Roberts, I., Greenwood, M. A. et al. (2 more authors) (2017) A Framework for Real-time Semantic Social Media Analysis. Journal of Web Semantics. ISSN 1570-8268

https://doi.org/10.1016/j.websem.2017.05.002

A Framework for Real-time Semantic Social Media Analysis

Diana Maynard, Ian Roberts, Mark A. Greenwood, Dominic Rout, Kalina Bontcheva

University of Sheffield
Department of Computer Science
Regent Court, 211 Portobello
S1 4DP, Sheffield, UK
d.maynard@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 three 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, an investigation of attitudes towards climate change expressed by these politicians and the public, via their engagement with environmental topics, and an analysis of public tweets leading up to the UK's referendum on leaving the EU (Brexit) in 2016. 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: Natural Language Processing, semantic search, social media analysis, Linked Open Data, semantic annotation, sentiment analysis

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 investigating people's values, and their opinions on specific topics such as the economy, rather than their opinions on particular parties as a whole, seems to give better insight.[2] Furthermore, simple sentiment analysis tools that look at people's opinions [1] often do not deal well with nuances such as sarcasm, nor the fact that people tend to express their sentiment about very specific events rather than about a party overall, which may have subtle differences. We therefore need much more sophisticated 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 topics (e.g. global warming, terrorism or immigration).

The unique nature of social media data is precisely what makes it also so challenging [2]. It is fast-growing,

[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

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. By this we mean 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 [3]. 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) about Scotland.

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 [2, 4, 5, 6, 7]. 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 [1] and SentiWordNet [8] and/or generic basic entity and topic recognition tools such as DBpedia Spotlight [9], or core open source NLP tools such as ANNIE [10] and Stanford CoreNLP [11], which are not adapted to the domain and task.

As a partial solution to these challenges, we present a semantic-based framework for real-time social media analysis, which combines a series of tools inside a flexible architecture that allows each component to be easily adapted to the specific social media monitoring task and its domain. For each application scenario, one simply selects the tools required for that task, which may be a combination of existing components and new ones specific to the task. There is thus no single system that can be installed; but rather, what is provided is an open-source toolkit of commonly used components, openly available web-based services, and a methodology for customising and combining these to the needs of each specific application.

The framework includes data collection, semantic analysis, aggregation, semantic search, and visualisation tools, which allow analysts to dig deep into the data and to perform complex semantic search queries over millions of social media posts, in near-real time. Furthermore, the semantic search and visualisation tools enable analysts to find new and interesting correlations between the data, a task which traditionally has been done manually and therefore on very small volumes of data. The paper includes a number of examples of semantic search and result visualisation for different applications, in order to demonstrate how the tool can be used by non-expert users (e.g. social scientists, political scientists, journalists) to get real-time insights into large-scale social media streams. The framework is highly scalable and can be used both for off-line processing and live processing of social media.

Semantic annotation and search are core to the framework, as they enable users to find information that is not based just on the presence of words, but also on their meaning [12]. First, automatically recognised entities and topics are disambiguated and linked to Open Data resources via URIs (e.g. DBpedia, GeoNames). Secondly, semantic knowledge from these resources is used to power semantic full-text search [13] over the social media stream. This kind of search draws both on document content and on semantic knowledge, in order to answer queries such as: "flooding in cities in the UK" or "flooding in places within 50 miles of Sheffield". In this case information about which cities are in the UK or within 50 miles of Sheffield is the result of ontology-based search (e.g. against DBpedia or GeoNames). Documents are then searched for the co-occurrence of the word "flooding" and the matching entities from the ontology-based search. In other words, what is being searched here is a combination of the document content for keywords, the index of semantically annotated entities that occur within these documents, and the formal knowledge.

The paper is structured as follows. First the generic framework and components are described in Section 2. Next, Section 3 shows 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

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

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 describe how the 2015 election application was adapted for the analysis of tweets about the EU Referendum in 2016 (Brexit) and give examples of some of the analysis performed. In Section 7 we present and discuss some evaluation of the analysis tools, and then conclude by discussing future directions.

2. An Open Source Framework for Social Media Analysis

The social media analytics toolkit is based around GATE [14], a widely used, open source framework for Natural Language Processing (NLP). The toolkit can perform all the steps in the analytics process: data 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.

2.1. Processing Overview

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 [13]. 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 re-run 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.

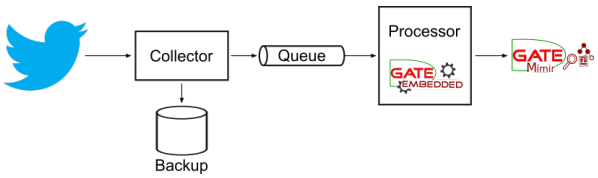

Figure 1: Simple architecture of live processing system

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,

[4] https://gate.ac.uk/gcp/

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 deduplication 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 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.

2.2. Semantic Annotation

GATE has recently been extended to provide numerous tools for social media analysis, namely automatic recognition of terms via TermRaider [15], named entities (people, places, organisations, dates etc.) via TwitIE [16], 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) [17, 18]. Where appropriate, entities and terms are associated with relevant URIs from Linked Open Data (LOD) via YODIE [19]. 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 LOD resources (e.g. DBpedia [20], GeoNames[5], GEMET[6], Reegle[7]), which are accessed via the GraphDB (formerly known as OWLIM) knowledge repository [21]. These are used both during semantic annotation and for semantic search and visualisations, as detailed next. The nature of the semantic annotation depends on the application: examples are given in the relevant sections of the paper. The purpose of the semantic annotation is to provide additional information that is not present in the documents themselves, but which can be used at query time to aggregate documents about the same concept/instance, or to get more specific information about a person, place or thing. For example, in the political scenarios, if one wants to know about the sentiment expressed by all politicians in Yorkshire, or about the proportion of posts which mention hospitals, or all tweets by MPs over the age of 50, this information is not explicit in the text but can be accessed by semantic annotation, as we shall explain.

The semantic information is acquired through various means: linking mentions of MPs and political parties to NUTS, DBpedia and YourNextMP; and via YODIE to link mentions of other persons, locations and organisations to their entries in DBpedia. In the environmental scenarios, mentions of environmental terms are extracted and linked to existing knowledge bases such as GEMET, Reegle and DBpedia, so that again, extra information is provided (for example, alternative names for the same term, hyponyms and hypernyms, and related information). This builds on previous work where LOD vocabularies were applied to semantic enrichment of environmental texts in the EnviLOD project [22]. A more detailed explanation of the general semantic annotation and querying process used here can be found in [23, 13].

2.3. 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* [24]. Semantic search requires some NLP techniques for understanding word meaning, typically Named Entity Recognition [25] and semantic annotation [26]. 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 indexed documents themselves, e.g. which political party a Member of Parliament (MP) belongs to, so that we can search for all documents written by or which mention MPs from a particular party. It also allows disambiguation of terms:

[5] `http://www.geonames.org`
[6] `https://www.eionet.europa.eu/gemet`
[7] `http://www.reegle.org`

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". Essentially, using semantic information as part of a search query allows us to retrieve documents that could never be found using any other search approach that relied purely on information from within the indexed documents.

After analysis, the social media posts are indexed using GATE Mímir [13], 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 representing, while the constituencies are in turn linked to higher-level geographic regions. This allows 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.

2.4. 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 [27].

Such information discovery and visualisation functionalities are provided by GATE Prospector [13], 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.

2.5. 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.

2.6. Availability

Most of the components of the framework described in this article are open source and freely available as part of GATE (under the LGPL licence). This includes not only the semantic annotation, indexing, and visualisation components, but also the GATE Cloud Paralleliser, which enables their scalable integration and execution. Prospector is not yet available as open-source as it is still under development and is currently difficult to configure. Also the visualizations shown in this article are application specific and not generally useful beyond this.

In order to help users with the configuration, adaptation, and use of these tools, we have also made them available via the GATE Cloud NLP-as-a-service platform[8]. With the components hosted on GATE Cloud users can easily configure their own Tweet collector, analysis pipeline (either a custom GATE application or one of the many existing applications that are available including TwitIE, YODIE, the DecarboNet Environment Annotator, and the Brexit Analyser), and either retrieve the annotated documents for further analysis or have them indexed within their own private GATE Mímir instance.

3. The Political Futures Tracker - Monitoring the UK 2015 Election

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 Members of Parliament (MPs) and parliamentary candidates (and responses to those tweets) throughout the 2015 UK 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.[9] 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.[10]

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[11] and all currently known election candidates[12] (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.

[8] `https://cloud.gate.ac.uk`

[9] `http://www.nesta.org.uk`

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

[11] 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/`

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

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 [16], identifies Persons, Places, Organisations etc., while **Named Entity Linking**, using YODIE [19], 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.[13] 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.[14] Hashtag preprocessing was added, in order to re-tokenise hashtags according to their constituent words [28]. 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 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 [18, 28], 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 useful 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 2.3, 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

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

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

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 is available for public use.[15]

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 topics mentioned in tweets by the MP or candidates for the Sheffield Hallam constituency, an example of a tweet found, and the semantic links that make the search possible. Neither the fact that the tweet author (Nick Clegg) is an MP, nor the details of which constituency he represents, are explicitly mentioned in the text; all that information comes from querying our extended DBpedia. We should note here that the query syntax 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

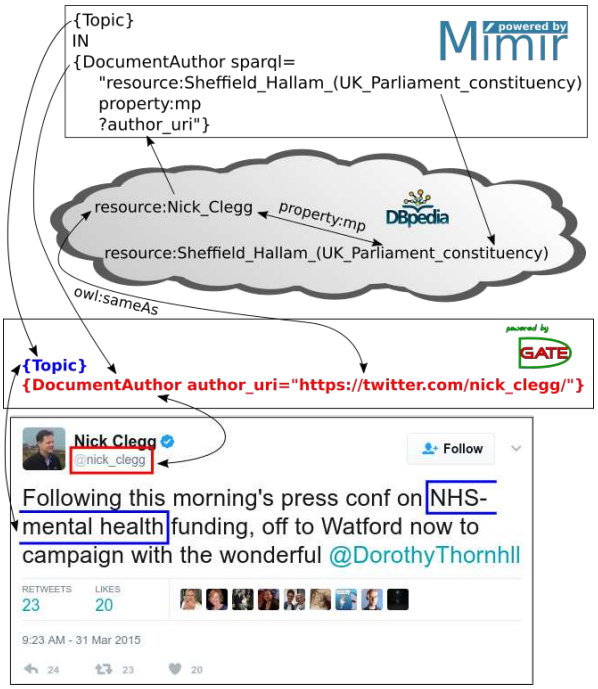

Figure 2: Example of a Mímir query, a matching tweet, and the semantic information that links them

of such a front-end for querying news can be seen at `http://demos.gate.ac.uk/pin/`.

4. Semantic Searches

This section describes how the framework was used to perform semantic 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 a 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 represen-

[15] `https://gist.github.com/greenwoodma/`

tative: Plaid Cymru (the Party of Wales), who have only 2 tweeting MPs, had an average of 110 tweets per MP, with the SNP (Scottish National Party) next highest at an average of 85.83 tweets (and 6 tweeting MPs).

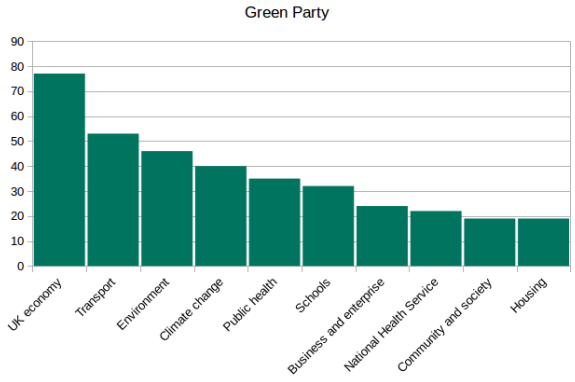

Figure 3: Top 10 topics mentioned by MPs from the Green Party

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. The following query shows a search for all tweets about the UK economy written by an MP from the Green Party; we then repeat this for each party and topic.

```
{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 shows the general purpose UI for exploring associations between semantic annotations and/or 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) are mapped to the two dimensions of a matrix, while the 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 by the 10 most frequent groups of politicians tweeting, where a group represents a political party and a category (MP or Candidate).[16]

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 2 March 2015. This is a dynamic visualisation, based on the Leaflet library[17] 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.[18]

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 shows the top 20 topics mentioned by MPs and candidates from the Sheffield Hallam constituency. The data shown is the result of the following Mímir semantic search query:

```
{Topic} IN {DocumentAuthor sparql=
"<http://dbpedia.org/resource/
Sheffield_Hallam_
(UK_Parliament_constituency)>
nesta:candidate|dbp-prop:mp ?author_uri"}
```

[16] "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.

[17] http://leafletjs.com/

[18] http://www.nesta.org.uk/blog/4-visualisations-uk-general-election

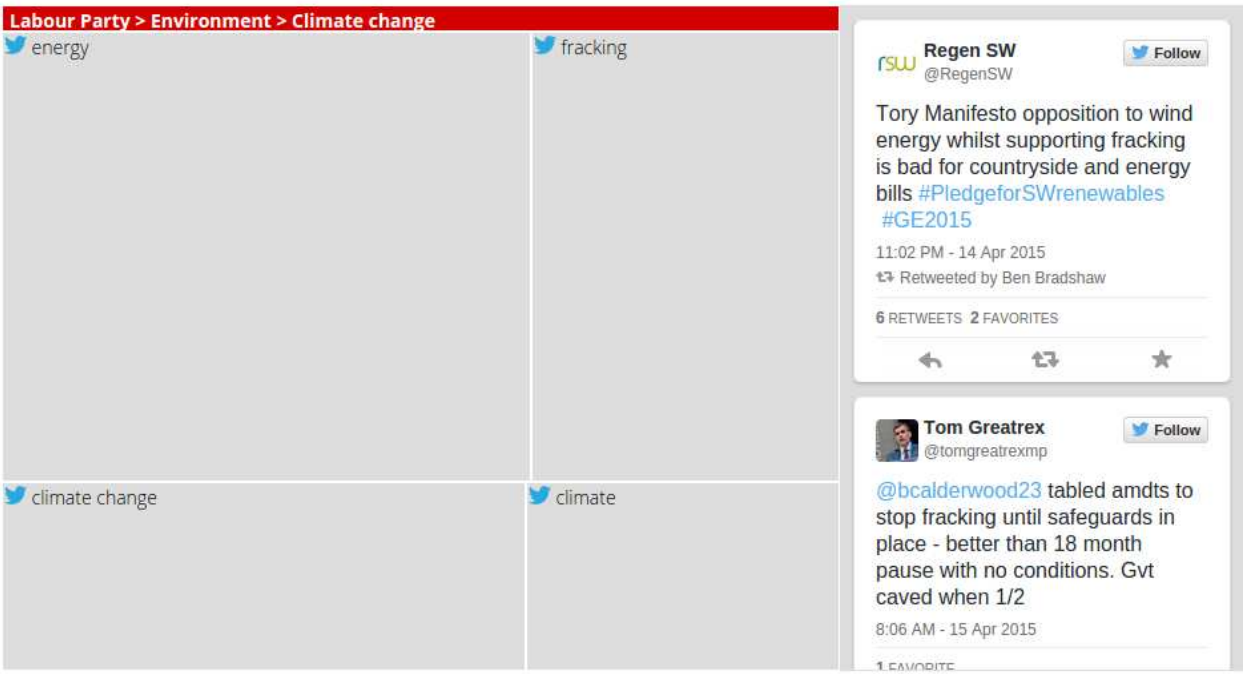

Figure 4: Treemap showing most frequent terms about climate change mentioned by the Labour Party

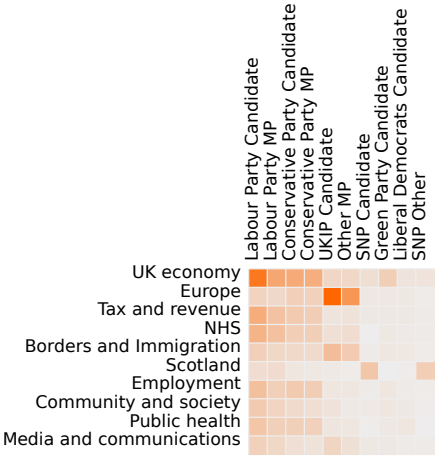

Figure 5: Prospector's Dynamic Co-occurrence Matrix

In essence this query finds all `Topic` annotations within documents where the `author_uri` feature contains a URI which appears in the result of the embedded SPARQL query fragment; full details of the Mímir query syntax is outside the scope of this article but can be found in [13]. The query fragment is expanded by Mímir before it is run against a SPARQL endpoint. The full query can be seen in Figure 8.

Prospector then builds frequency and co-occurrence statistics over the selected tweets for the selected semantic annotation type (`Topic` 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.

This example illustrates that we can use semantic information to help select document subsets for further processing, utilizing information not explicitly contained within the documents themselves. In this instance this information was that the author of the tweet was the MP or a candidate for the Sheffield Hallam constituency, but it could easily have been any other related semantic information, such as authors born in a certain location, authors educated at a specific University, documents containing mentions of locations within a given constituency, etc.

5. Measuring Climate Change Engagement

In our second (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

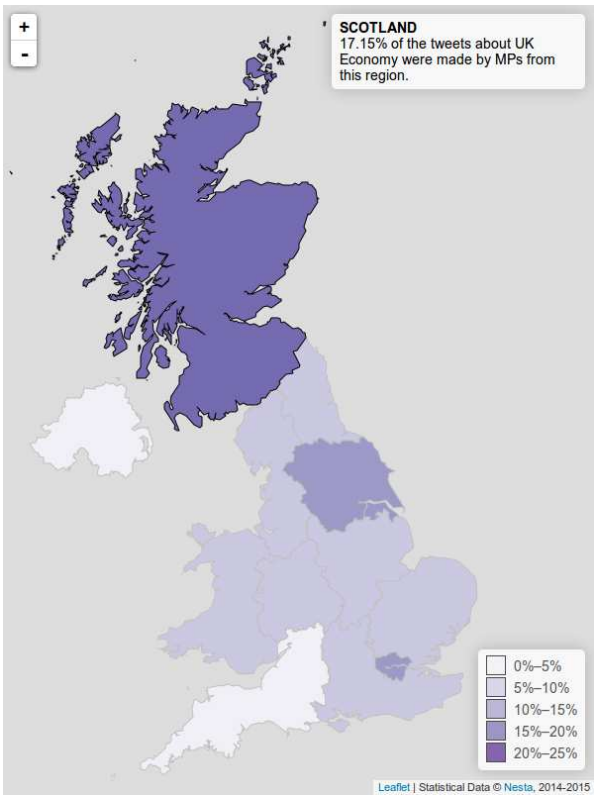

Figure 6: Choropleth depicting distribution of tweets about the economy

change issues is still problematic. The EU DecarboNet project[19] 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 [29]. 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, comparing it with other political topics.

We measured engagement with the different political topics described in Section 3 in four ways. First, we looked at retweets. We found a high number of climate change related retweets, which typically indicates a high level of engagement [7]. 64.48% of the climate change tweets in our dataset were retweets, and 94.3% of them were either retweets or replies. The percentage was much higher than for many other topics such as schools (57% retweets, and 90% retweets and replies).

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

Second, we looked at sentiment, which has previously been shown to be a good indicator of engagement [6]. Figure 9 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 [6]. Again, "climate change" scored 3rd highest (after "business and enterprise" and "schools"). Finally, we investigated the number of URLs found in climate change-related tweets. In Boyd's study of random tweets [30], 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 (UK Independence Party) 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. Analysis of Brexit tweets

Our third case study, the real-time Brexit monitor, was developed to analyse tweets relating to the 2016 EU membership referendum in the UK, as they came

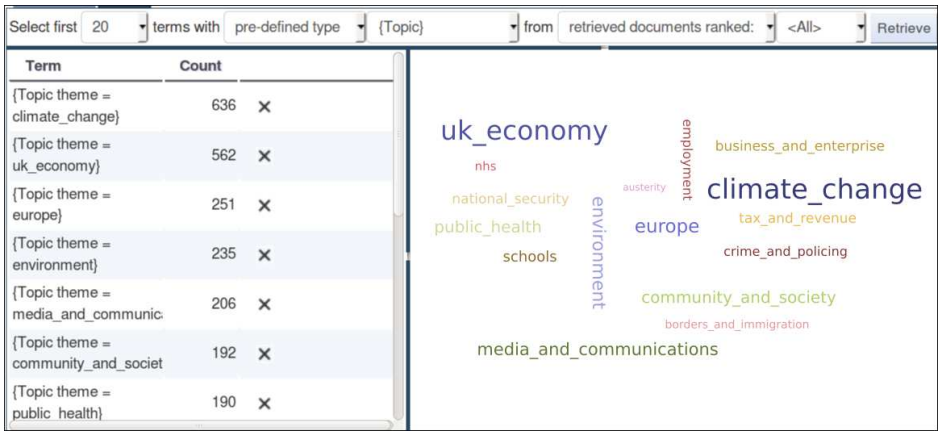

Figure 7: GATE Prospector showing the Topics mentioned by MPs and candidates from the Sheffield Hallam constituency

```
PREFIX :<http://dbpedia.org/ontology/>
PREFIX dbp-prop:<http://dbpedia.org/property/>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>
PREFIX owl:<http://www.w3.org/2002/07/owl#>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX nesta:<http://nesta.org.uk/property/>
PREFIX twitter:<https://twitter.com/>
PREFIX ynmp:<https://yournextmp.com/person/>

SELECT DISTINCT ?author_uri WHERE {
<http://dbpedia.org/resource/Sheffield_Hallam_(
↪  UK_Parliament_constituency)> nesta:candidate|dbp-prop:mp ?
↪  author_uri
}
```

Figure 8: Fully Expanded SPARQL Query

in, in order to track the debate unfolding on Twitter. Unlike other Brexit analysis tools, the aim was not to try to predict the outcome of the referendum nor to answer the question of whether Twitter can be used as a substitute for opinion polls. Instead, our focus was on a more in-depth analysis of the referendum debate; the people and organisations who engage in those debates; what topics were discussed and opinions expressed, and who the top influencers were.

As with the Political Futures Tracker, the Brexit monitor analysed and indexed tweets in real time, in order to identify commonly discussed topics and opinions expressed. It also examined specifically whether a tweet was expressing support for remaining in or leaving the EU.

The analysis tools consisted of TwitIE, theme and topic detection, and topic-centric sentiment analysis, as used in the Political Futures Tracker. The topic detection and sentiment analysis tools were adapted to deal better with certain terms relevant to Brexit. We then added a Leave/Remain classifier, described in Section 6.2, which helped us to identify a reliable sample of tweets with unambiguous stance. Finally, we added a tweet geolocation component, which used latitude/longitude, region, and user location metadata to geolocate tweets within the UK NUTS2 regions. The architecture is depicted in Figure 10.

6.1. Statistical analysis

The tweets were collected based on a number of referendum-related hashtags and keywords, such as #voteremain, #voteleave, #brexit, #eureferendum. On average, the number of original tweets, replies, and retweets was close to half a million per day, with 60% of them retweets. On referendum day itself, we had to analyse in real-time well over 2 million tweets, which averaged just over 23 tweets per second. Tweet volume picked up dramatically as soon as the polls closed at 10pm, and we were consistently getting around 50 tweets per second and were also being rate-limited by the Twitter API. Interestingly, amongst the 1.9 million tweets collected in the first 4 days, only 134,000 contained a URL (7%). Amongst the 1.1 million retweets, 11% contained a URL, which indicates that tweets with URLs tend to be retweeted more. This is in line with theories of social media engagement [6]. These low percentages suggest that the majority of tweets on the EU referendum were expressing opinions or addressing another user, rather than sharing information or providing external evidence. Although a heavy volume of tweets was published, we can see that with only 6.8% of these being replies, and over 58% retweets, the debate on Twitter resembles an echo chamber.

6.2. Hashtags as a predictor of Leave/Remain support

One question we were interested in answering was how reliable hashtags would be as a predictor of a tweet supporting either the Leave or Remain stance. Over

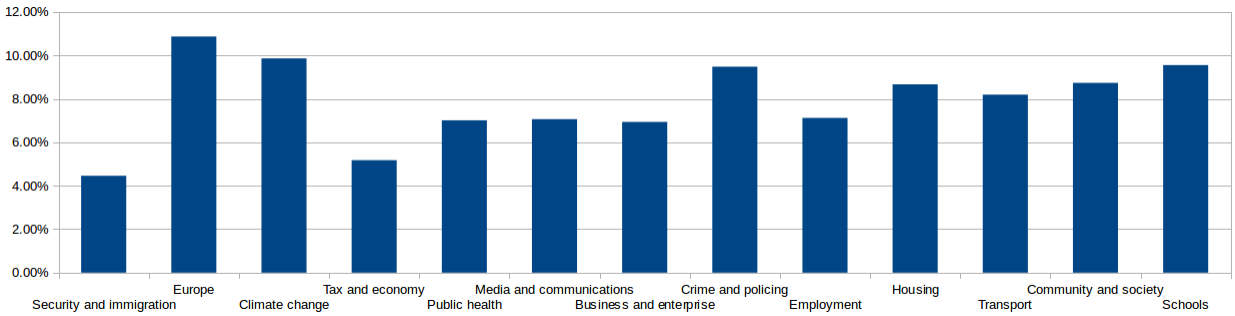

Figure 9: Percentage of opinion-bearing tweets per topic

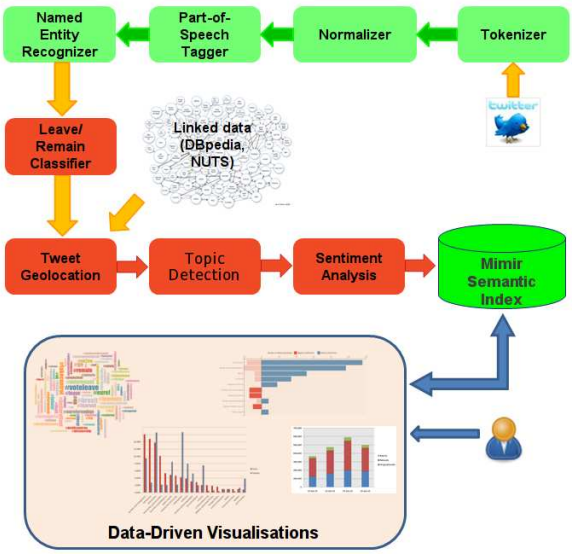

Figure 10: Architecture of the Brexit monitor

56% of all tweets on the referendum contained at least one hashtag. Some of these were clearly indicative of support for the leave/remain campaigns, e.g. #votetoleave, #voteout, #saferin, #strongertogether. There were also hashtags which tried to address undecided voters, e.g. #InOrOut, #undecided.

A recent study by Ontotext[20] classified EU referendum tweets as leave or remain using a set of 30 hashtags, based on whether each tweet contained predominantly leave or remain hashtags. Through manual examination of a random sample, however, we found that this strategy does not always deliver a reliable assessment, since in many cases leave hashtags are used as a reference to the leave campaign, while the tweet itself is supportive of remain or neutral (and similarly for remain hashtags). A more reliable though slightly more restrictive approach is to consider the last hashtag in the tweet as the most indicative of its intended stance. This results in a higher precision sample of remain/leave tweets, which can then be analysed in more depth in terms of topics discussed and opinions expressed. We are currently crowdsourcing 5,000 human annotated tweets with final hashtags, so the accuracy of the different hashtag heuristics can be measured more reliably.

Using this approach, amongst the 1.9 million tweets between June 13th and 19th, 5.5% (106,000) were identified as supporting the Leave campaign, while 4% (80,000) as supporting the Remain campaign. Taken together, this constitutes just under a 10% sample, which we considered sufficient for the purposes of our analysis.

6.3. Analysis of voting trends

We performed a number of different analyses of the tweets, too numerous to describe here. One of the most interesting was the analysis of voting trends. Separating the tweets into original tweets, replies, and retweets, we applied our Leave/Remain classifier to all tweets posted on or after 1pm on June 22nd, but before voting closed at 10pm on June 23rd. On this set, we identified 39,000 advocating Remain and 61,000 for Leave. On June 23rd, as Twitter activity picked up significantly, we found 291,000 matching tweets. Unlike other studies, however, our voting intent heuristic identified 164,000 tweets advocating Leave and only 127,000 advocating Remain. While voting tweets from @Brndstr and tweet volume statistics from #EURef Data Hub both indicated that Remain was dominant, this trend was not supported in our voting intention sample.

7. Evaluation

While the analysis toolkit has many interesting features and can provide valuable insights into social media

[20] http://ontotext.com/twitter-users-support-brexit/

(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 [31, 32]. 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 [16], and performed better than two state-of-the-art Twitter-specific systems, Stanford-twitter [33] and a tool developed by Ritter [34], achieving 80% F1-measure on a corpus of tweets. In a more recent comparison of 5 state-of-the-art NER tools for Twitter [35], TwitIE outperformed the others in terms of Organisation recognition, although it did worse on Person and Location recognition (probably due to lack of relevant training data and smaller gazetteers, but possibly also due to different annotation guidelines for these types).

The Named Entity Linking and Disambiguation component YODIE was evaluated in [36] against two state-of-the-art tools DBpedia Spotlight [9] and Zemanta[21], achieving the highest Precision (67.59%) and 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.

We have recently compared an improved version of YODIE for tweets on an unseen test set of 191 tweets from the corpus described in [37]. This corpus comprises 794 tweets of which approximately half come from the field of finance and online news, and half from tweets about climate change. The training set from this corpus was used together with the AIDA training, TAC 2009, TAC 2011 and TAC2012 corpora for training the candidate selection model. The test set contains 191 tweets with a total of 5100 tokens, of which 3333 are word tokens, and with 225 linkable (non-NIL) target annotations. YODIE was compared against 5 other state-of-the-art tools: AIDA 2014 [38]; DBpedia Spotlight; TagMe [39]; TextRazor[22]; and Zemanta. The experimental setup is described in detail in [40]. Table 1 shows the results; we can see that YODIE outperforms all the other tools in terms of F-measure, although Aida has a better Precision and TagMe has better recall. While the scores are all quite low, they clearly show improvement on the previous version, and obtain the best results when combining Precision and Recall. Work is still ongoing to improve performance further.

System	Precision	Recall	F1.0
YODIE	0.50	0.61	**0.55**
Aida 2014	**0.59**	0.38	0.46
Spotlight	0.09	0.51	0.15
TagMe	0.10	**0.67**	0.17
TextRazor	0.19	0.44	0.26
Zemanta	0.48	0.56	0.52

Table 1: Evaluation of YODIE on tweets

Note that we expect some further improvements in the performance of YODIE in the coming months. This is due to two reasons. First, we have detected some target annotations in the gold standard corpus which are not quite correct and need to be fixed. Second, the model is trained on the previous release of DBpedia, while the application itself uses the newer version, so the features are slightly different. We plan to retrain the model on the latest release of DBpedia. The addition of YODIE in the toolkit, even if not perfect, enables us to enrich the search. In the political use case, NUTS is used for location information (e.g. which area does a constituency belong to), but it is only linked to consituencies. If we want to query any of our datasets for things like all hospitals in a region, or all towns in Yorkshire, this would be impossible without the addition of YODIE, though as noted above, it can only be as good as its data sources.

An earlier version of the environmental term recognition component has been evaluated in [41] and showed promising results. On a corpus of climate change tweets, it achieved Precision of 81.49%, Recall of 82.82% and F1 of 82.15%. We expect the results on the political dataset to be higher because since that evaluation we have improved the Recall considerably by

[21] Originally available at `http://www.zemanta.com`, but no longer exists.

[22] `http://www.textrazor.com`

adding the term expansion techniques. On that corpus, TwitIE scored a Precision of 85.87% for Named Entity Recognition, 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 [42]. On a corpus of environmental tweets, it achieved accuracy of 86.80%, beating three other state-of-the-art systems DIVINE [43], ARCOMEM [31] and SentiStrength [1]. 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.

8. Related Work

The main purpose of our framework is to provide a methodology and practical toolkit for analysing high-volume social media content. There are two main elements to this: first, the tools for the data analysis; and second, the querying and visualisation aspect. Both of these are critical to the success of the system: without an in-depth data analysis, the insights one can draw will be limited; and without easy ways to query and visualise the data, the accessibility of this for non-expert users is limited. The main novelty of our work lies in the combination of all these elements (including also scalability, flexibility, adaptability and availability) in a single framework.

One of the major challenges in the 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 British 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 involves 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 [44] and tweets [45, 46]. For instance, Phelan *et al* [47] use word clouds to present the results of a Twitter based recommendation system. The Eddi system [48] 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 [46]. More recently, the concept has been extended towards image clouds [49].

The main drawback of cloud-based visualisations is their static nature. Therefore, they are often combined with timelines showing keyword/topic frequencies over time [50, 48, 51, 52], as well as methods for discovery of unusual popularity bursts [44]. [53] 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 [54] 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 [49], falling keyword bars [51] and dynamic information landscapes [51]).

In addition, some systems try to capture the semantic relatedness between topics in the media streams. For instance, BlogScope [44] 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 [51]. Topic-document relationships can be shown also through force-directed, graph-based visualisations [55]. Lastly, Archambault *et al* [56] 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 [51]) 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 [50], lists of tweets [54], and mood maps [50]. Aggregated sentiment is typically presented using pie charts [52] and, in the case of TwitInfo, the overall statistics are normalised for recall

[54]).

Other tools which try to analyse and visually represent the information in document collections are not specifically aimed at social media and do not capture information such as sentiment. These are more aimed at general semantic annotation (mapping information in the text to relevant ontology classes, and finding relational information). Work by [57] analysed documents about the Dutch elections, and also included semantic technologies, with a search and visualisation tool similar to Mímir, but this had much less functionality overall than our framework (for example, no ability to use SPARQL queries). The NeBro visualisation tool they used does not seem to exist any more.

A number of recent research initiatives have focused on the use of sentiment and social media analysis to understand citizens' opinions about governments and governmental agencies. [58] used a topic modelling approach, via an enhanced form of tf.idf, to understand what topics were being discussed by the public on social media, and what were the root causes. They combine the topics with some pre-defined keywords for each topic and a basic sentiment lexicon. However, they do not use any semantic technologies and thus their analysis is limited, just showing some sentiment about various topics. Similarly, [59] propose the use of semantic role labelling to detect the semantic arguments in tweets and understand the political sentiment of citizens, while [60] use semantic role labelling to help resolve the data sparsity problem in tweets by clustering similar tweets based on their content. All these techniques help to enrich the tweets with further semantic information, but do not make use of external information or semantic web technologies such as Linked Open Data.

Most existing social media analysis tools tend to use shallow textual and frequency-based information. The contribution of our work lies in a deep analysis of the meaning of the text, 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 underpins the data aggregation (e.g. location-based, party-based) and visualisation UIs. This means that one can query at a much more insightful level than traditional analysis tools, as evidenced in our use cases. 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 [61].

In summary, previous approaches to the social and semantic analysis of data such as political tweets typically do not combine all the different kinds of information both within and external to the tweets, nor do they provide such rich functionality for analysing the data (i.e. the combination of full-text, linguistic and semantic queries), in a manner which can also be easily adapted to new tasks and domains.

9. Conclusions

This paper has 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 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 at both a general analysis of the political discourse in the run up to the 2015 UK general elections, and 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 was extended to cover the House of Commons debates, which included more information about the political roles MPs fulfil. The aim of this was to investigate the evolution of context in an organizational setting, looking at indicators such as changes to ontologies over time [62].

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 men-

tioned. 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.

Finally, the Brexit monitor demonstrates how the tools can easily be adapted to a new scenario. While still in the politics domain, the tasks here were a little bit different, such as extending the opinion mining tools to deal specifically with stance detection. Extensions were added to the original components, and new kinds of questions were investigated using the semantic search and visualisation tools. Our techniques for investigating stance detection also enabled us to get a fresh insight on the voice of the community, something which more simple analysis tools failed to pick up. Post-hoc analysis of this dataset is still ongoing.

With respect to the framework itself, future work will focus on widening the kinds of semantic annotation services within, to include better coverage of languages other than English. In addition, data collection and processing of other kinds of social media content will be added, e.g. Reddit, Facebook, Instagram. We also plan on extending the GATE Cloud user interface with the ability to customise the semantic annotation components via web-based user interfaces. More corpus-level statistics will also be offered, as well as network-based visualisations of the social media analysis (e.g. @mention graphs).

10. Acknowledgments

This work was partially supported by the European Union under grant agreements No. 610829 DecarboNet and 654024 SoBigData, the UK Engineering and Physical Sciences Research Council (grant EP/I004327/1), and by the Nesta-funded Political Futures Tracker project (http://www.nesta.org.uk/news/political-futures-tracker).

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

[2] K. Bontcheva, D. Rout, Making sense of social media through semantics: A survey, Semantic Web - Interoperability, Usability, Applicability 5 (5) (2014) 373—403.

[3] S. Wibberley, J. Reffin, D. Weir, Method51 for mining insight from social media datasets, COLING 2014 (2014) 115.

[4] K. Holmberg, I. Hellsten, Analyzing the climate change debate on twitter content and differences between genders, in: Proceedings of the ACM WebScience conference, Bloomington, IN, USA, 2014, pp. 287–288.

[5] R. Pfitzner, A. Garas, F. Schweitzer, Emotional divergence influences information spreading in twitter., ICWSM 12 (2012) 2–5.

[6] C. Meili, R. Hess, M. Fernandez, G. Burel, Earth hour report, Tech. Rep. D6.2.1, DecarboNet Project Deliverable (2014).

[7] M. Rowe, H. Alani, Mining and comparing engagement dynamics across multiple social media platforms, in: Proceedings of the 2014 ACM conference on Web science, ACM, 2014, pp. 229–238.

[8] A. Esuli, F. Sebastiani, SentiWordNet: A publicly available lexical resource for opinion mining, in: Proceedings of LREC 2006, 2006.

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

[10] H. Cunningham, D. Maynard, K. Bontcheva, 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, Association for Computational Linguistics, Stroudsburg, PA, USA, 2002, pp. 168–175. doi:10.3115/1073083.1073112.
URL http://gate.ac.uk/sale/acl02/acl-main.pdf

[11] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, D. McClosky, The Stanford CoreNLP natural language processing toolkit, in: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2014, pp. 55–60.
URL http://www.aclweb.org/anthology/P/P14/P14-5010

[12] A. Kiryakov, B. Popov, D. Ognyanoff, D. Manov, A. Kirilov, M. Goranov, Semantic annotation, indexing and retrieval, Journal of Web Semantics 1 (2) (2004) 671–680.

[13] V. Tablan, K. Bontcheva, I. Roberts, H. Cunningham, Mímir: an open-source semantic search framework for interactive information seeking and discovery, Journal of Web Semantics 30 (2015) 52–68.
URL http://dx.doi.org/10.1016/j.websem.2014.10.002

[14] H. Cunningham, V. Tablan, A. Roberts, K. Bontcheva, Getting more out of biomedical documents with GATE's full lifecycle open source text analytics, PLoS Computational Biology 9 (2) (2013) e1002854. doi:10.1371/journal.pcbi.1002854.
URL http://dx.doi.org/10.1371/journal.pcbi.1002854

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

[16] K. Bontcheva, L. Derczynski, A. Funk, M. A. Greenwood, D. Maynard, 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.

[17] D. Maynard, G. Gossen, M. Fisichella, A. Funk, Should I care

about your opinion? Detection of opinion interestingness and dynamics in social media, Journal of Future Internet.

[18] D. Maynard, K. Bontcheva, 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.

[19] G. Gorrell, J. Petrak, K. Bontcheva, G. Emerson, T. Declerck, Multilingual resources and evaluation of knowledge modelling - v2, Tech. Rep. D2.3.2, Trendminer Project Deliverable (2014). URL http://www.trendminer-project.eu/images/d2.3.2_final.pdf

[20] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, 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 (2009) 154–165.

[21] 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.
URL http://www.ontotext.com/sites/default/files/publications/Kiryakov_OWLIM_www2006.pdf

[22] K. Bontcheva, J. Kieniewicz, S. Andrews, M. Wallis, Semantic Enrichment and Search: A Case Study on Environmental Science Literature, D-Lib Magazine 21 (1/2).

[23] D. Maynard, K. Bontcheva, I. Augenstein, Natural Language Processing for the Semantic Web, Morgan and Claypool, 2016.

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

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

[26] K. Bontcheva, H. Cunningham, Semantic annotations and retrieval: Manual, semiautomatic, and automatic generation, in: J. Domingue, D. Fensel, J. Hendler (Eds.), Handbook of Semantic Web Technologies, Springer, 2011, pp. 77–116.

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

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

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

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

[31] D. Maynard, Challenges in Analysing Social Media., in: A. Duşa, D. Nelle, G. Stock, G. G. Wagner (Eds.), Facing the Future: European Research Infrastructures for the Humanities and Social Sciences, SCIVERO Verlag, Berlin, 2014.

[32] L. Derczynski, D. Maynard, G. Rizzo, M. van Erp, G. Gorrell, R. Troncy, K. Bontcheva, Analysis of named entity recognition and linking for tweets, Information Processing and Management 51 (2015) 32–49. doi:10.1016/j.ipm.2014.10.006.

[33] J. Finkel, T. Grenager, 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, Association for Computational Linguistics, 2005, pp. 363–370.

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

[35] K. Greenfield, R. Caceres, M. Coury, K. Geyer, Y. Gwon, J. Matterer, A. Mensch, C. Sahin, O. Simek, A reverse approach to named entity extraction and linking in microposts, in: 6th workshop on 'Making Sense of Microposts', 2016, pp. 67–69.

[36] L. Derczynski, D. Maynard, N. Aswani, 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.

[37] G. Gorrell, J. Petrak, K. Bontcheva, Using @Twitter conventions to improve #lod-based named entity disambiguation, in: The Semantic Web. Latest Advances and New Domains, Springer, 2015, pp. 171–186.

[38] J. Hoffart, Y. Altun, G. Weikum, Discovering emerging entities with ambiguous names, in: Proceedings of the 23rd International Conference on World Wide Web, WWW '14, 2014, pp. 385–396.

[39] P. Ferragina, U. Scaiella, Tagme: On-the-fly annotation of short text fragments (by wikipedia entities), in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10, ACM, New York, NY, USA, 2010, pp. 1625–1628. doi:10.1145/1871437.1871689.
URL http://doi.acm.org/10.1145/1871437.1871689

[40] D. Maynard, J. Petrak, A. Funk, L. Derczynski, Multilingual content processing methods, Tech. Rep. D3.1, COMRADES Project Deliverable (2016).

[41] D. Maynard, A. Scharl, Text analytics tools for environmental information extraction v.2, Tech. Rep. D2.2.2, DecarboNet Project Deliverable (2016).

[42] D. Maynard, K. Bontcheva, Challenges of Evaluating Sentiment Analysis Tools on Social Media, in: Proceedings of LREC 2016, Portoroz, Slovenia, 2016.

[43] S. Gindl, A. Weichselbraun, A. Scharl, Cross-domain contextualisation of sentiment lexicons, in: Proceedings of 19th European Conference on Artificial Intelligence (ECAI-2010), 2010, pp. 771–776.

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

[45] D. Shamma, L. Kennedy, E. Churchill, Tweetgeist: Can the Twitter timeline reveal the structure of broadcast events?, in: Proceedings of CSCW 2010, 2010.
URL http://research.yahoo.com/pub/3041

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

[47] O. Phelan, K. McCarthy, B. Smyth, Using Twitter to recommend real-time topical news, in: Proceedings of the 2009 ACM Conference on Recommender Systems, 2009, pp. 385–388.

[48] M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, 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), 2010, pp. 303–312.

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

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

[51] A. Hubmann-Haidvogel, A. M. P. Brasoveanu, A. Scharl, M. Sabou, S. Gindl, Visualizing contextual and dynamic features of micropost streams, in: Proceedings of the #MSM2012

Workshop, CEUR, Vol. 838, 2012.

[52] J. Y. Weng, C. L. Yang, B. N. Chen, Y. K. Wang, S. D. Lin, IMASS: An Intelligent Microblog Analysis and Summarization System, in: Proceedings of the ACL-HLT 2011 System Demonstrations, Portland, Oregon, 2011, pp. 133–138.
URL http://www.aclweb.org/anthology/P11-4023

[53] N. Diakopoulos, M. Naaman, 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, 2010, pp. 115–122.

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

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

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

[57] W. Van Atteveldt, S. Schlobach, F. Van Harmelen, Media, politics and the semantic web, in: European Semantic Web Conference, Springer, 2007, pp. 205–219.

[58] R. Arunachalam, S. Sarkar, The new eye of government: Citizen sentiment analysis in social media, in: Sixth International Joint Conference on Natural Language Processing, 2013, p. 23.

[59] S. S. Hasbullah, D. Maynard, Automated Content Analysis: A Sentiment Analysis on Malaysian Government Social Media, in: ACM IMCOM (International Conference on Ubiquitous Information Management and Communication), Danang, Vietnam, 2016.

[60] X. Liu, K. Li, M. Zhou, Z. Xiong, Collective semantic role labeling for tweets with clustering, in: IJCAI, Vol. 11, Citeseer, 2011, pp. 1832–1837.

[61] B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, M. Goranov, KIM – Semantic Annotation Platform, in: 2nd International Semantic Web Conference (ISWC2003), Springer, Berlin, 2003, pp. 484–499.

[62] M. A. Greenwood, V. Solachidis, O. Papadopoulou, K. Apostoliditis, D. Galanopoulos, D. Tastzoglou, V. Mezaris, B. Eldesouky, N. K. Tran, C. Hube, C. Niederée, J. Petrak, G. Gorrell, ForgetIT Deliverable D6.4: Contextualisation Framework and Evaluation (February 2016).