29th International Conference on Informatics for Environmental Protection (EnviroInfo 2015)
Third International Conference on ICT for Sustainability (ICT4S 2015)

Understanding climate change tweets: an open source toolkit for social media analysis

Diana Maynard
Dept. of Computer Science
University of Sheffield
Sheffield, S1 4DP
Email: d.maynard@sheffield.ac.uk

Kalina Bontcheva
Dept. of Computer Science
University of Sheffield
Sheffield, S1 4DP
Email: k.bontcheva@sheffield.ac.uk

***Abstract*—Collective awareness about climate change is an ongoing problem because there is such a wealth of information available, which can be confusing, contradictory and difficult to interpret. In order to help citizens understand environmental concerns, and to help organisations better inform and target interested people with campaigns, we have developed an open source toolkit to analyse social media data on the topic of climate change. The toolkit comprises methods for extracting, aggregating, and visualising actionable knowledge, based on automatic analysis of large volumes of text. The key terms, topics and sentiments expressed in online discussions are extracted, along with key indicators of climate change, and are stored in a semantic search tool, which enables complex searches over the huge volumes of data. We describe a scenario using the toolkit to gain insights from a large collection of political tweets, showing how we can analyse this dataset for understanding engagement of the public with respect to the topic of climate change.**

I. Introduction

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 information, making it hard to know what is accurate and relevant. On the one hand, ways to make sense of all this information are necessary, while on the other hand, policy makers and experts need better methods to inform people. The EU Decarbonet project[1] aims to close this Information - Action - Behaviour loop via social innovation principles, helping to empower citizens by means of tailored information services, with the ultimate goal of improving collective awareness.

This paper presents research from the DecarboNet project on an open-source toolkit, comprising methods for extracting, aggregating, and visualising actionable knowledge, based on automatic analysis of large volumes of social media content. The key terms, topics and sentiments expressed in online discussions on climate change are extracted, along with key indicators of climate change, and are stored in a semantic search tool [1], which enables complex searches over the huge volumes of data. We also describe a practical example of how these open source tools were used to gain insights from a large collection of political tweets, showing how we can analyse this dataset for understanding engagement of the public with respect to the topic of climate change and the environment. This is important not just for promoting best practices in both individual and collective climate change mitigation, but also for understanding and influencing the complex interaction between climate change and politics. In communities where political apathy is rife, climate change is nevertheless one topic in which people seem to engage more readily with politics, in order to instigate the changes they believe are necessary [2].

The development of this toolkit for social media analysis was motivated by the challenges posed by the need to analyse large volumes of social media [3]. Microposts such as tweets are, in particular, extremely challenging to analyse at scale, especially when opinions are concerned, since the genre is noisy; tweets have little context and assume much implicit knowledge; and utterances are often short. As such, conventional Natural Language Processing (NLP) tools typically do not perform well when faced with tweets [4], and their performance also negatively affects subsequent analysis, search, and visualisation.

Ambiguity is a particular problem for tweets, since we cannot easily make use of extended contextual information: unlike comments in blog posts, tweets do not typically follow a conversation thread, and are analysed independently of each other, since they arrive as a continuous information stream. They also exhibit much more language variation, and make frequent use of slang, emoticons, abbreviations and hashtags, which can form an important part of the meaning. Typically, they also contain extensive use of irony and sarcasm, which are particularly difficult for a machine to detect [5]. On the other hand, their terseness can also be beneficial in focusing the topics more explicitly: it is fairly rare for a single tweet to be related to more than one main topic, which can thus aid disambiguation by emphasising situational relatedness. In longer user-generated posts such as blogs, comments on news articles and so on, a further challenge is raised by the tracking of changing and conflicting interpretations in discussion threads [6].

Our open-source toolkit for social media analytics helps with addressing these challenges by offering an extensible and modular set of text analysis, aggregation, and search components, tailored specifically to analysing social media at scale and in near real-time.

II. Related Work

Information Extraction (IE) [7], [8] is a form of automatic text analysis, which extracts fixed-type, unambiguous snippets

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

as output. The extracted data may be used directly for display to users (e.g. a list of named entities mentioned in a document), for storing in a database for later analysis, or for improving search and other information access tasks.

Named Entity Recognition (NER) is one of the key information extraction tasks, which is concerned with identifying mentions of names of entities such as people, locations, organisations and products. It is typically broken down into two main phases: *entity detection* and *entity typing* (also called classification). A follow-up step to NER is Named Entity Linking (NEL), which links entities mentioned within the same document (also known as co-reference), or in other resources, such as Linked Open Data (also known as entity resolution). Typically, state-of-the-art NER and NEL systems are developed and evaluated on news articles and other carefully written, longer content [9], [10].

Information extraction from social media, and microblogs in particular, has recently become an active research topic [11], following early experiments which showed this genre to be extremely challenging for state-of-the-art algorithms [12]. For instance, named entity recognition methods typically have 85-90% accuracy on longer texts, but only 30-50% on tweets [13], [14]. To combat these problems, research has focused on microblog-specific information extraction algorithms (e.g. named entity recognition for Twitter using CRFs [13] or hybrid methods [15]). Particular attention is given to microtext normalisation [16], as a way of removing some of the linguistic noise prior to part-of-speech tagging and entity recognition. However, as described in [17], there are many challenges still to be overcome in the analysis of social media. Furthermore, tools for language analysis often need to adapted specifically to the domain in order to get best results; little work has been done specifically on applying such tools to the climate change domain.

There has also been some recent work on semantic analysis of environmental science documents, but there are still many outstanding challenges. Most research has focused on geospatial information [18], with applications including GIS environments/Spatial Data infrastructures (SDI), environmental sensor networks and geotagging [19]. These approaches all identify interdisciplinary datasets, and apply semantic enrichment in order to improve search and enable correct use of data [20]. The LOD GEMET thesaurus underpins the EU INSPIRE directive, which aims to establish a digital infrastructure for spatial information in Europe in order to support environmental research, policy and decision-making.

Linked Open Data (LOD) vocabularies have been applied to semantic enrichment of environmental science literature in the EnviLOD project [21], on which we build here. However, the complex and irregular nature of social media, as described above, means that existing text analysis tools for this domain (e.g. those from the EnviLOD project) are not suitable to be applied directly on tweets. Moreover, the focus in EnviLOD was on identifying and disambiguating geo-locations, rather than on the recognition of environmental terms and indicators.

The work presented here on engagement analysis on climate change-related tweets aims to complement and extend the prior research of Meili *et al* [22], which focused specifically on tweets about Earth Hour. Other similar analyses have been performed on social media data. For example, Cheong and Lee [23] studied the impact of social media on Earth Hour 2009 and found a direct correlation between high social media activism and reduced energy usage. Rowe and Alani [24] studied the role of engagement dynamics across different social media platforms, defining features which were indicative of strong social media engagement; we have used some of these in our experiments to measure political engagement, as discussed in Section VIII. The idea of political engagement is discussed from a more philsophical point of view in a number of research studies; for example, Stoker [25] explains the phenomenon of disengagement as being linked to a global disenchantment with governmental processes, parties and the whole political system. Our hypothesis that climate change goes against this trend is supported, on the other hand, by the fact that it is a topic that society believes it can actively do something about, without relying solely on the government. The theroetical underpinnings of this are discussed in detail in [2].

Most existing social media search and visualisation methods tend to use shallow textual and frequency-based information. One of the main contributions of our work lies in taking into account the extra semantic knowledge about the entities, terms, and sentiment mentioned in the messages, based on information from Linked Open Data resources such as DBpedia, and in utilising deeper linguistic information that aims at true understanding of the meaning behind the text. This semantic knowledge also underpins the data aggregation and visualisation UIs shown in the application scenario. In addition, our framework enables further 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 [26].

III. The Open Toolkit for Large-Scale, Environmental Social Media Analysis

The toolkit comprises a number of text analysis components, which are aimed specifically at social media content (see Figure 1). Each component can be used on its own, as a stand-alone web service or can be combined together with the rest, to form a sequential pipeline. The underlying architecture which supports such flexible component-based design is GATE [27], a widely used, open source framework for text analysis. For scalability and real-time analysis over large streaming data, the GATE Cloud paralleliser was used [28].

Low-level linguistic pre-processing of the social media content, such as tokenisation, normalisation, and part-of-speech tagging, is carried out by reusing the TwitIE plugin from GATE [4]. In a nutshell, TwitIE [4] performs low-level linguistic analysis, as well as named entity recognition (NER) over tweets. It was developed especially to handle the noisy and idiosyncratic nature of tweets. It contains components for language identification; a specialised tokeniser for handling emoticons, user names, URLs, hashtags, etc.; a part-of-speech tagger trained on tweets; a text normalisation component to handle typical slang and abbreviations; and a tweet-customised set of named entity recognition rules.

The newly-created components of the social media toolkit (see Figure 1) address the automatic recognition of environmental terms, named entities (people, places, organisations,

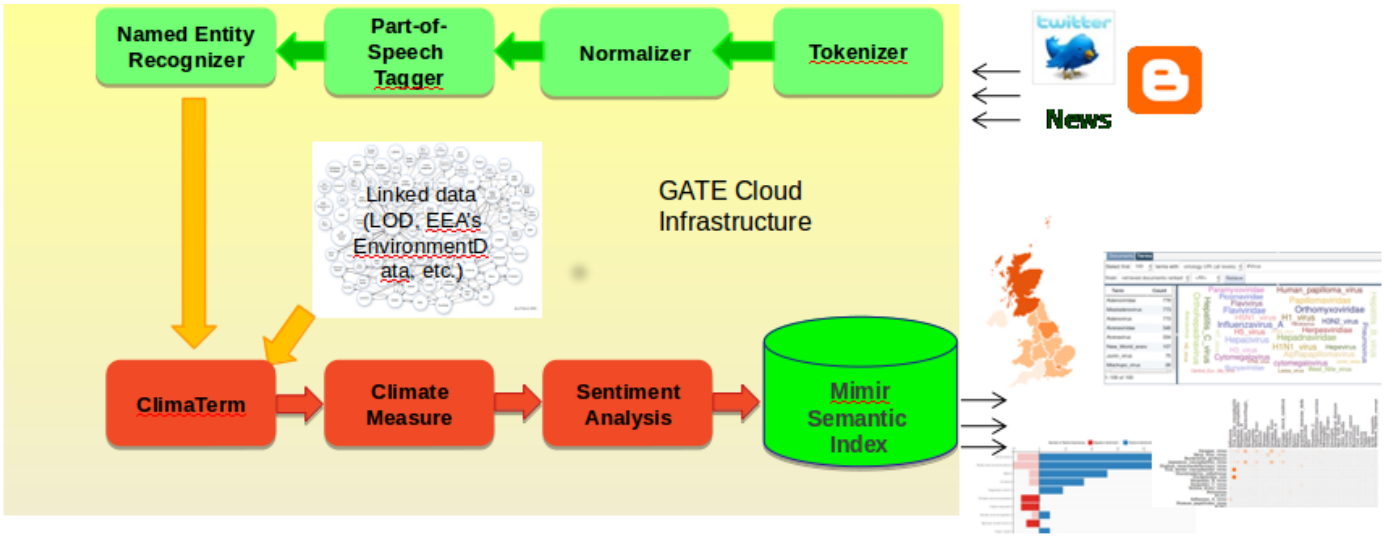

Fig. 1. Toolkit Architecture Diagram: Red components are newly developed and climate-change oriented; Green components are generic and reused from GATE

dates etc.), topics (general themes, e.g. "the EU", "climate change", "economy" and so on), and 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). Where appropriate, entities and terms are associated with relevant URIs from Linked Open Data.

The following sections describe each of these in more detail. Firstly, Section IV introduces the named entity recognition and term extraction components which constitute the ClimaTerm application, and some experiments to test its performance. Next, the ClimateMeasure component recognises automatically mentions of environmental indicators (Section V) and makes use of ClimaTerm to do so. The sentiment detection component is presented in Section VI. Ultimately, all these tools are combined together, in order to analyse and index a large corpus of political tweets, over which users can perform semantic search, as explained in Section VII.

IV. Identifying Environmental Terms

The ClimaTerm component annotates mentions of terms related to climate change and the environment. It runs as a standalone web service[2], as well as within GATE as part of the integrated social media analysis framework described in Section III. The web service takes as input a document or set of documents, and outputs those documents as XML files annotated with term and URI information. The underlying application consists of the following processing stages:

- (reused) linguistic pre-processing via TwitIE: tokenisation, sentence splitting, part-of-speech tagging, morphological analysis, named entity tagging;
- (newly developed) term extraction: matching against known terms, plus some recognition of morphological and synonym variants
- (reused) export as XML

[2]available at http://services.gate.ac.uk/decarbonet/term-recognition

A. Environmental Term Recognition

The term extraction component is based on lexical matching against two environmental ontologies: GEMET[3] and Reegle[4], and then enhancing these with linguistic rules to gain more coverage and handle morphological variants.

GEMET (GEneral Multilingual Environmental Thesaurus) is the reference vocabulary of the European Environment Agency (EEA) and its Network (Eionet). It was conceived as a "general" thesaurus, aiming to define a core of general terminology for the environment, and contains 5208 terms originating from a number of different thesauri. From this, we extracted all the terms along with their label and URI, as in the example entry below:

```
label=air pollutant
URI=http://www.eionet.europa.eu/gemet/
concept/263
```

The Reegle clean energy and climate glossary contains 2527 terms related to climate change in RDF format and a SPARQL endpoint. We extracted the URI, prefLabel, and scopeNote information from this ontology, as shown in the example entry below:

```
prefLabel= crop yield increase
URI= http://reegle.info/glossary/1400
scopeNote= how and where yields might
increase due to climate change.
```

We also extracted additional 965 terms listed as "alternative label" to the main terms. For example, "wind power frequency changers" is the alternative label for the term "windpower inverters". In most cases, these are synonymous or close-to-synonymous terms.

We observed that some of the entries in GEMET and Reegle are, in fact, named entities (mainly names of organisations, such as "World Wildlife Fund"). Since these are already

[3]http://www.eionet.europa.eu/gemet/
[4]http://www.reegle.info/glossary

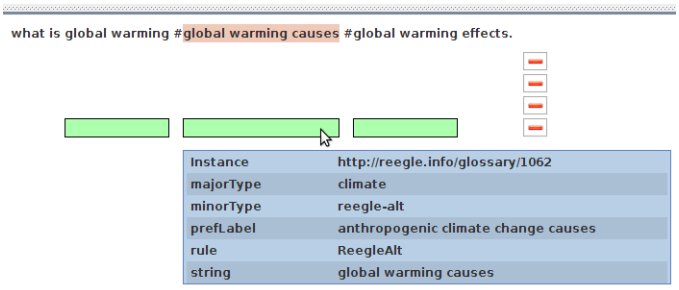

Fig. 2. Annotation of a term variant in GATE

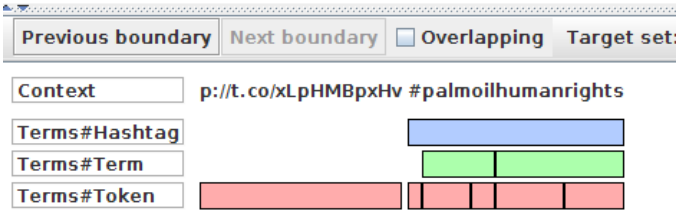

Fig. 3. Hashtag decomposition in GATE

recognised by TwitIE, they were excluded from consideration during term recognition, in order to avoid duplicate annotations.

Figure 2 shows the term "global warming causes" found in a tweet and annotated with respect to the Reegle glossary. The green boxes show the various terms found in that tweet (e.g. "global warming"). The features depicted in the popup window (in blue) give the URI of the term (instance), the type of term (climate-related, and coming from the altLabel property in the ontology), the preferred label of that term (prefLabel), the rule fired (for debugging purposes) and the original string.

A number of reasons caused terms to be missed by matching solely against these ontologies. Some frequently occurring terms were missing from the ontologies, which we discovered through corpus analysis and added manually as a list of other important terms, related to climate change and the environment. Second, a large number of missing terms were due to hashtags where a multiword term was combined into a single word and was therefore not recognised, for example #palmoil. Other missing terms included morphological variants of multi-word terms. To deal with these issues, some new pre-processing components had to be added, as follows.

First, the frequently occurring missing terms were discovered using the TermRaider term extraction tool[5]. This is not used on its own for the automatic extraction of climate change terms, because initial experimentation showed that it could not differentiate climate change-specific terms from more general terms. Instead, we ran it over a large corpus of climate change-related tweets and extracted the top 250 terms, then manually analysed this list and added any missing environmental and climate change terms, which were not already covered by GEMET and REEGLE, e.g. "permafrost".

Next, hashtag pre-processing was added, in order to re-tokenise hashtags according to their constituent words [5]. This enables, for example, the term "palm oil" to be matched against the text "#palmoil", as depicted in the screenshot in Figure 3. The figure shows the span of the original hashtag (in blue, and denoted by the row Terms#Hashtag'), the terms found within the hashtag (in green, and denoted by the row "Terms#Term") and the new tokens (in red, and denoted by the row "Terms#Token"). The original hashtag "#palmoilhumanrights" has been correctly tokenised into four words, and then two terms have been found, each consisting of two words ("palm oil" and "human rights"). Without the retokenisation, correct term identification would have been impossible.

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

Finally, linguistic restrictions and rules were added, in order to improve coverage and accuracy. One such rule is that terms which are not part of noun phrases should not be annotated. For example "global warming causes", where "causes" is a plural noun, could be a relevant term, but if "causes" is a verb, then it is not part of the term. Compare the two sentences "Global warming causes stratospheric cooling." and "Global warming causes are still not entirely without debate." This does bring some additional issues, however, since the TwitIE POS tagging is not perfect, and also due to homographs such as "lead" (which could be a verb or noun), but initial experiments show it to be a worthwhile tradeoff.

B. *ClimaTerm Evaluation*

Evaluation experiments were performed to check the validity of extracted terms, using three different corpora, all collected using DecarboNet's Media Watch on Climate Change platform[6]. These were then annotated manually by a student and verified by one of the researchers.

First we compare the performance of the tool using only GEMET, only Reegle and the combined ontologies on these datasets, measuring Precision, Recall and F-measure. Table I shows the results on our human-annotated corpus of 455 climate-related tweets; performance was similar on the other two corpora. As can be seen from the results, the best performance is obtained with the combined set; however recall still needs to be improved, which is the focus of future work.

Term set	P	R	F1
Gemet	84.96	45.80	59.51
reegle	95.72	31.55	47.46
Combined	85.87	53.05	65.58

TABLE I. EVALUATION OF DIFFERENT TERM SETS ON CLIMATE CORPUS

In order to evaluate the benefit from having social media-specific linguistic pre-processing, we compared ClimaTerm's performance on term recognition when two different linguistic pre-processing pipelines were used: ANNIE, the standard GATE set for longer documents, vs the social media-optimised TwitIE. Best overall results were achieved with TwitIE (65.8% F1 vs 60.82%), although precision was slightly higher with ANNIE.

The human annotators were asked not just to annotate all mentions of environmental and climate change terms, but also to express how confident they were of their own judgement (high, medium, low). Of the original 1523 terms in the tweet

[6]http://www.ecoresearch.net/climate

corpus, 1154 were marked with high confidence, 320 terms with medium and 40 with low confidence. Therefore, our next experiment was to examine how confidence impacts term recognition performance. The best results were obtained when ClimaTerm used only the high and medium-ranked terms (see Table II). As expected, recall increased but precision decreased as lower quality terms were included. However, removing the low confidence terms from the set did not improve performance significantly, so the remaining evaluations used the full set of terms.

Term set	P	R	F1
H	72.93	58.15	64.71
H+M	86.09	53.73	66.17
H+M+L	85.87	53.05	65.58

TABLE II. EVALUATION OF DIFFERENT TERM SETS ON CLIMATE CORPUS

Finally, we compared the annotations found in GEMET and Reegle against those found by a general purpose term recognition component (TermRaider), which performs single and multi-word term recognition based on tf.idf and other statistical measures. TermRaider performance on the manually annotated gold standard gave precision of 45.64%, recall of 74.49% and an F1 of 56.60%. In this case, precision is quite low, because TermRaider finds many terms that are not domain-specific. Table III shows some examples of terms found only in TermRaider, Reegle and GEMET respectively.

TermRaider only	GEMET only	reegle only
Arctic biodiversity	agriculture	sustainability
abrupt climate change	deforestation	anthropogenic climate change
renewable energy	Antarctica	geothermal
evolution	biofuel	biodiesel
shark	ecology	palm oil industry

TABLE III. TERMS UNIQUE TO EACH TERMSET

V. IDENTIFYING CLIMATE CHANGE INDICATORS

The second tool we present, ClimateMeasure[7], is a web service which complements the ClimaTerm service described above, by identifying indications of the presence of a quantitative measurement related to climate change. For example, this might include changes in mortality rates for a country or population, percentage decrease in forest areas and so on.

ClimateMeasure extracts useful indicators of climate change such as "energy use", "carbon pollution", etc. for particular locations, together with measurable effects such as percentages and measurements. It uses a manually compiled list of indicator seed terms, plus the TwitIE linguistic preprocessing tools; annotations of measurements and percentages (created by the GATE Measurements component) and domain-specific terms (from ClimaTerm). These components are described in more detail next.

The application checks for the presence of an indicator, a location, and a measurement or percentage in the tweet. Indicators are identified via list lookup. An initial list of 42 seed terms was defined manually, followed by an iterative process to retrieve further terms. Relevant tweets were extracted from the MWCC dataset, using the seed terms as keywords (to ensure relevance), and then processed with TermRaider. We used TermRaider rather than ClimaTerm because we wanted to ensure maximum recall. Precision was not important in this case, since the list of top ranked, newly discovered indicators was post-edited and filtered manually by one of the researchers. The process can be repeated as necessary to find further new environmental indicator terms.

[7]available at http://services.gate.ac.uk/decarbonet/indicators/

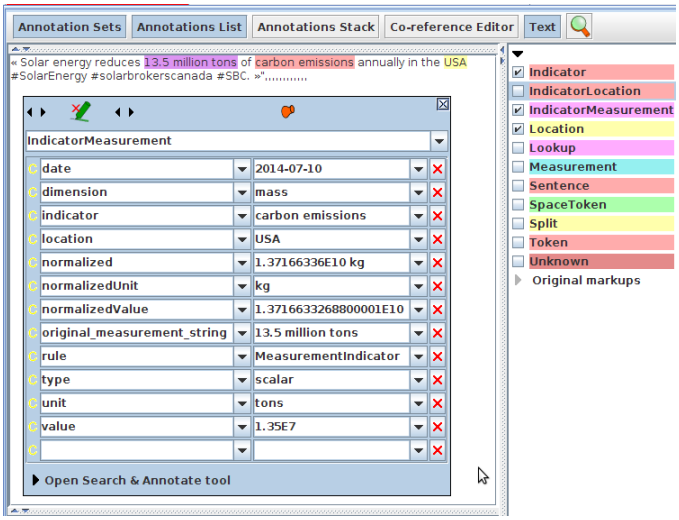

Fig. 4. Annotation of normalised measurement indicator in GATE

Locations are found by the following method. They are first identified in the body of the text or in hashtags via TwitIE. If no Location is found, the usernames are checked to see if a location is contained there (e.g. USAToday, age_uk. Failing this, the tweet metadata is checked for the presence of a value of target_location and this is used instead. We use the tweet metadata as the last resort, because it is quite unreliable (since it only depicts the location of the person tweeting, but may not be relevant to the subject matter of the tweet).

Measurements and percentages are identified using the GATE Measurements component and via TwitIE respectively. The measurements are also normalised to their SI unit, so that the same measurements in different systems (e.g. acres and square metres) can be equated. This normalisation process is described more fully in [29]. Examples of the output of the application are shown in Figure 4, which depicts a measurement with its normalisation information.

Finally, dates are identified by the following process. First, the body of the tweet is checked for mention of a date, using TwitIE and this is then normalised. Date normalisation ensures that all dates are represented in the same format (DD-MM-YYYY) and that relative dates are represented as absolute dates with respect to today's date. For example, if today is 1 September 2014, a mention of "tomorrow" in the text will be represented as 02-09-2014. A feature is also added to indicate whether the date is in the past, present or future. If this fails to retrieve a date, the tweet metadata is checked for the presence of a date. Again, this is a last resort, because it is not necessarily very accurate – the date of the tweet might not be the same date as the subject matter of the tweet.

While this application has not yet been formally evaluated, manual inspection of the results shows that there are a few

cases where indicators are being missed, largely due to more complex grammatical structures being used (for instance, coordinations of amounts and dates are not always correctly dealt with). Other errors are rarer, but generally due to inaccuracies from the TwitIE pre-processing, e .g. if a Location is wrongly identified or missing. In general, the results are of high quality, however.

VI. Sentiment Detection

The sentiment detection component aims to identify the kinds of opinions being expressed towards climate change and environmental topics or entities. This is useful for a number of reasons. Organisations need to have a better understanding of public perception of climate change, in order to develop campaigns and strategies: automatic opinion mining can help them to understand what are the opinions on crucial topics and events. It is also useful to know how these opinions are distributed in relation to demographic user data, how they evolve over time, who are the opinion leaders, and what is their impact and influence. This kind of information helps to improve both the development and marketing of environment-related tools and technology, by better understanding social perception and behaviour. Currently, it is hard and time-consuming to get this information by traditional means, such as youGov polls[8]. Existing tools for sentiment analysis are often not tailored to the domain and also fail to understand issues such as slang and sarcasm typically found in social media. For example, many tools would wrongly classify the following tweet as positive: *"@adambation Try reading this article, it looks like it would be really helpful and not obvious at all."*.

The DecarboNet sentiment detection component is based on an adaptation of our core rule-based sentiment analysis tools [5], [30]. Although Machine Learning applications are more typically used for sentiment analysis tasks [31], [32], [33], there are advantages to using a rule-based approach in this situation. Firstly, there is no need for large amounts of training data, which are unavailable in this domain and also hard and time-consuming to create. Secondly, DecarboNet needs to cover multiple languages and we can re-use many components easily. Lastly, the system can more easily be adapted to different types of text (we work with both social media and longer, more formal kinds of text).

For DecarboNet we adapted our lexicons of positive and negative words to the climate change domain. General purpose lexicons were reused, such as swear words, emoticons, and sarcastic indicators. These are combined with a set of rules, to determine the nature and strength of the sentiment (positive or negative), who the opinion holder is, what topic the sentiment refers to, and whether it is sarcastic or not. The rules for sentiment strength and score combine a number of linguistic features such as adverbs, negation, conditional sentences, questions, swear words, sarcasm indicators and so on. Further rules then attempt to link the correct opinion holder and topic with the sentiments found (opinion-target matching), by extracting sentiment-containing words in a linguistic relation with terms/entities. For example, in the phrase "life flourishing in Antarctica", we would annotate "flourishing" as a positive sentimen word, and "Antarctica" as a Location, and then connect the two. Currently we use a farly simple notion of linguistic relation, using shallow parsing techniques, because more complex parsers do not work well on noisy, ungrammatical data such as tweets. The relation matching essentially operates by chopping the tweet or sentence into phrases and preferring closest matches within phrases, but considers also the confines of any linguistic constraints such as conditionals. Figure 5 shows an example of an automatically annotated sarcastic tweet. We can see that the sentiment word identified was "nice" but that because sarcasm was detected, the polarity was reversed from positive to negative. Note that in this case, no particular target was found for the sentiment, so it is just recorded as a general negative tweet.

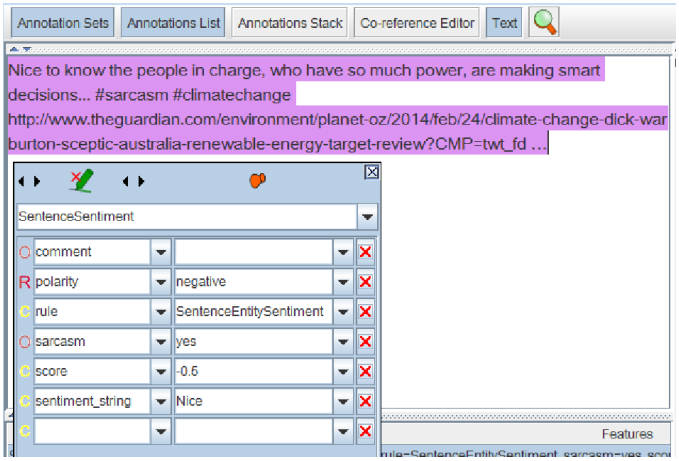

Fig. 5. Annotation of a sarcastic tweet

VII. Semantic Search

After automatic analysis, social media content is indexed using GATE Mimir [1], which enables complex semantic searches to be performed over the entire dataset. Semantic search over documents is about finding information that is based not just on the presence of words, but also on their meaning [34], [1]. GATE Mimir is an integrated semantic search framework, which offers indexing and search over full text, document structure, document metadata, linguistic annotations, and any linked, external semantic knowledge bases. It supports hybrid queries that arbitrarily mix full-text, structural, linguistic and semantic constraints.

The benefit of GATE Mimir, semantic search, and the grounding of automatically discovered information into ontologies is that we can search not just over things that are explicitly mentioned in the text (e.g. specific terms), but also for implicit information, based on knowledge in the ontologies. Cambridge, for example (as well as many other names and words), has multiple meanings, i.e. is ambiguous. The token "Cambridge" may refer to the city of Cambridge in the UK, to Cambridge in Massachusetts, the University of Cambridge, etc. Similarly, different tokens may have the same meaning, e.g. New York and the Big Apple. Therefore, semantic search tries to offer users more precise and relevant results, by using knowledge encoded in ontologies.

As shown in the following section, Mimir allows us to search for environmental terms expressed in a multitude of

[8]https://yougov.co.uk/

different ways (thanks to the results from the environmental text analysis components), including synonyms and hypernyms of the terms mentioned. We can search for not just a particular politician saying something about climate change (which may be expressed in a number of ways and need not mention climate change specifically), but for any Labour MP, based on knowledge about UK MPs which is encoded formally in DBpedia [35]. Similarly, we can search not just for a particular climate term such as "solar energy" but for all terms related to the concept of energy (even if they do not contain the actual word "energy", such as "electricity"). The knowledge that electricity and energy are related comes from GEMET or Reegle and does not need to be explicit in the documents. Furthermore, the analysis is not limited to searching for relevant documents that match a query, but we can also answer more complex questions such as "Which political party talks the most about environmental topics?", "Which politician gets the most retweets when they talk about climate change?", or "In which area of the country are people most engaged in climate change topics on social media?"

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 [36].

Such information discovery and visualisation functionalities are provided in our toolkit by GATE Prospector [1], which includes visualisation of correlations, frequency statistics, map-based plots, and timeseries-based visualisations. 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 document. We can also easily perform 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. The following section demonstrates how the complete social media analysis toolkit was used in a real-world scenario, to better understand climate change as a political topic, and to examine the public's engagment with it.

VIII. Measuring climate change engagement on political tweets

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. As a test case for our social media analysis toolkit, we experimented with a large number of tweets, in order to examine the extent to which climate change is resulting in more engaged citizens, when compared to other topics.

Our test case focuses on the interaction between UK members of parliament (MPs), election candidates and members of the public on Twitter. As part of the Political Futures Tracker[9], a large corpus of political tweets was collected and analysed in real-time using Twitter's streaming API. The collection we are using here consists of all tweets by former UK MPs or known election candidates, and every retweet and reply to these (by any member of the public), between 24 October 2014 and 13 February 2015. This comprised approximately 1.8 million tweets, of which approximately 100k are original tweets, 700k are replies, and 1 million are retweets. Since the dataset itself will be growing continuously until after the UK elections on 7 May 2015, ultimately a much larger set of analysis results will be obtained.

We ran our new social media analysis toolkit on this streaming data, analysing and indexing it for search in real time. The tweets were annotated automatically with mentions of names of MPs, election candidates, and political parties; sentiment and opinions; other named entities; and high-level political topics (e.g. the domain term "fossil fuels" is an indicator of the "environment" topic). The list of high-level topics was derived from those used to categorise documents on the gov.uk website[10]. In cases where multiple topics were mentioned, these were each connected, where appropriate, to their respective related sentiment using linguistic principles. However, this is relatively rare in tweets which are typically short and are targeted towards a single main assertion. Tweets related to multiple politicians are rarely an issue, since we identify specifically the politician who was the author of the sentiment expressed in the tweet, or the author of the tweet. If a politician tweets about the opinion expressed by another politician, this is also captured.

The automated analysis made use of ClimaTerm, the DecarboNet sentiment analysis component, as well as specific tools for MP and election candidate disambiguation developed for the Political Futures Tracker (PFT). The analyzed data was then indexed using GATE Mimir, enabling us to perform complex queries and visualisations over the data. An example of such a visualisation is shown in Figure 6, which depicts a treemap enabling users to investigate the major topics and subtopics mentioned by the UK Labour Party. In the example, we have drilled down to look at what subtopics of climate change are mentioned in the 1.8 million tweets. Examples of tweets about fracking are shown. The size of the grey boxes denotes the proportion of mentions of each subtopic, within the climate change high-level topic.

For the engagement experiments, the corpus was divided into 12 topics, including climate change, immigration, Europe and employment (a subset of the original 42 topics used in the PFT). Engagement was measured by looking at the average number of retweets and replies per original tweet, the number of mentions of other users, the number of URLs mentioned, and the proportion of opinionated tweets, in particular those with positive sentiment. 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 with

[9]A project funded by Nesta: see https://gate.ac.uk/projects/pft/ for more details

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

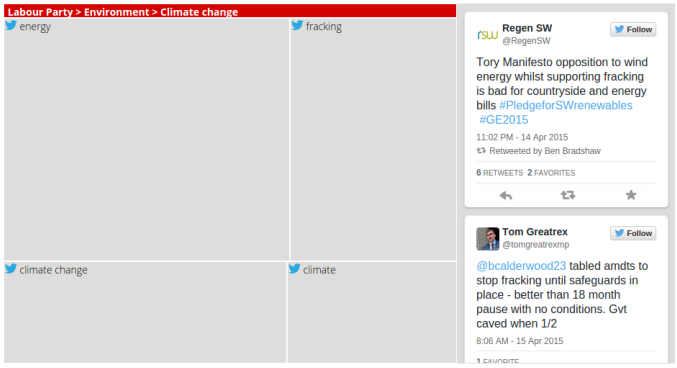

Fig. 6. Treemap showing tweets about fracking by the Labour party

climate change still ranks very highly, mostly residing in the top three of most engaged topics.

We found a large number of climate change related retweets, which indicates a high level of engagement. 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). Figure 7 shows the average number of retweets per original tweet for all topics, with climate change having the third highest score, after security and immigration, and Europe.

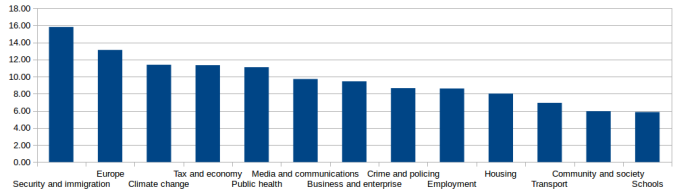

Fig. 7. Average retweets per original tweet

Looking at sentiment, which has been previously shown to be a good indicator of engagement [24], we found that climate change tweets were the second highest scoring topic, 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 with 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 [22]. Again, climate change scored 3rd highest (after "business and enterprise" and "schools"). Finally, we investigated the number of URLs found in climate change tweets, since it has been shown that original tweets containing a URL are more likely to be retweeted [37] and thus show engagement. In Boyd's study of random tweets, 52% of retweets contained a URL, and this figure is widely accepted to be the average. 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 a URL.

Although climate change still has a slightly lower engagement rate than topics such as Europe and the economy, which are hot topics in the buildup to the UK elections, and even though climate change is not frequently mentioned by most UK political parties in their tweets, engagement with climate change still ranks very highly, mostly residing in the top three of most engaged topics. Interestingly, climate change tweets contain the highest proportion of URLs compared with other topics, which 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 course, engagement can also be indicated by other factors: number of retweets and sentiment alone is not a complete indicator. In future, we could consider also investigating issues such as the strength of sentiment expressed, use of sarcasm and other specific linguistic features, for example. We also did not consider in this work the reputation or trustworthiness of the person tweeting [38], but this is clearly an important factor and has been left for future work.

IX. Conclusions

This paper has presented an overview of the open-source toolkit for analysing social media and its application to the analysis of a large volume of political tweets to measure engagement around climate change. The toolkit contains a number of different components for analysis, search and visualisation, which can be adapted to the domain and task, as demonstrated. We have shown here how it has been used to understand and track public engagement on climate change, and some experiments comparing climate change engagement with other political topics. This is only one example of using the toolkit, but is a clear demonstration of how the deeper forms of analysis can be used to understand online discourse on complex phenomena.

Future work will continue to extend the coverage of the environmental term recognition component in particular, and to perform more experiments on climate change data. It will also look at how the climate change indicators can be used for in-depth analysis of social media.

Acknowledgments

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

References

[1] 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. [Online]. Available: http://dx.doi.org/10.1016/j.websem.2014.10.002

[2] 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.

[3] D. Maynard, "Challenges in Analysing Social Media." in *Facing the Future: European Research Infrastructures for the Humanities and Social Sciences*, G. S. Adrian Dua, Dietrich Nelle and G. G. Wagner, Eds. Berlin: SCIVERO Verlag, 2014.

[4] 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.

[5] 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.

[6] S. Somasundaran and J. Wiebe, "Recognizing stances in ideological on-line debates," in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Association for Computational Linguistics, 2010, pp. 116–124.

[7] C. Cardie, "Empirical Methods in Information Extraction," *AI Magazine*, vol. 18, no. 4, 1997.

[8] D. E. Appelt, "An Introduction to Information Extraction," *Artificial Intelligence Communications*, vol. 12, no. 3, pp. 161–172, 1999.

[9] L. Ratinov and 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.

[10] D. Rao, P. McNamee, and M. Dredze, "Entity linking: Finding extracted entities in a knowledge base," in *Multi-source, Multi-lingual Inf. Extraction and Summarization*. Springer, 2013.

[11] M. Rowe, M. Stankovic, A. Dadzie, B. Nunes, and A. Cano, "Making sense of microposts (#msm2013): Big things come in small packages," in *Proceedings of the WWW Conference - Workshops*, 2013.

[12] 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.

[13] 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.

[14] X. Liu, S. Zhang, F. Wei, and M. Zhou, "Recognizing named entities in tweets," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 359–367.

[15] M. van Erp, G. Rizzo, and R. Troncy, "Learning with the Web: Spotting Named Entities on the intersection of NERD and Machine Learning," in *Proceedings of the 3^{rd} Workshop on Making Sense of Microposts (#MSM2013)*, 2013.

[16] B. Han and T. Baldwin, "Lexical normalisation of short text messages: makn sens a #twitter," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, ser. HLT '11, 2011, pp. 368–378.

[17] 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*, vol. 51, pp. 32–49, 2015.

[18] K. Janowicz, A. Bröring, C. Stasch, S. Schade, T. Everding, and A. Llaves, "A restful proxy and data model for linked sensor data," *International Journal of Digital Earth*, vol. 6, no. 3, pp. 233–254, 2013.

[19] W. Pillman, S. Schade, and P. Smits, *Innovations in sharing environmental observations and information, Proceedings of the 25th EnviroInfo Conference*. Shaker-Verlag, 2011.

[20] H. Schentz, J. Peterseil, B. Magagna, and M. Mirtil, "Semantics in ecosystems research and monitoring," in *Proceedings of the 25th International EnviroInfo Conference*, W. Pillman, S. Schade, and P.Smits, Eds., 2011.

[21] K. Bontcheva, J. Kieniewicz, S. Andrews, and M. Wallis, "Semantic Enrichment and Search: A Case Study on Environmental Science Literature," *D-Lib Magazine*, vol. 21, no. 1/2, 2015.

[22] C. Meili, R. Hess, M. Fernandez, and G. Burel, "Earth hour report," DecarboNet Project Deliverable, Tech. Rep. D6.2.1, 2014.

[23] M. Cheong and V. Lee, "Twittering for earth: A study on the impact of microblogging activism on earth hour 2009 in australia," in *Intelligent Information and Database Systems*. Springer, 2010, pp. 114–123.

[24] 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*. ACM, 2014, pp. 229–238.

[25] G. Stoker, "Explaining political disenchantment: finding pathways to democratic renewal," *The Political Quarterly*, vol. 77, no. 2, pp. 184–194, 2006.

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

[27] 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*, vol. 9, no. 2, p. e1002854, 02 2013. [Online]. Available: http://dx.doi.org/10.1371/journal.pcbi.1002854

[28] V. Tablan, I. Roberts, H. Cunningham, and K. Bontcheva, "GATECloud.net: a Platform for Large-Scale, Open-Source Text Processing on the Cloud," *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences*, vol. 371, no. 1983, p. 20120071, 2013. [Online]. Available: http://dx.doi.org/10.1098/rsta.2012.0071

[29] H. Cunningham, V. Tablan, I. Roberts, M. A. Greenwood, and N. Aswani, "Information Extraction and Semantic Annotation for Multi-Paradigm Information Management," in *Current Challenges in Patent Information Retrieval*, ser. The Information Retrieval Series, M. Lupu, K. Mayer, J. Tait, and A. J. Trippe, Eds. Springer Berlin Heidelberg, 2011, vol. 29, pp. 307–327.

[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*, in press.

[31] E. Boiy and M.-F. Moens, "A machine learning approach to sentiment analysis in multilingual web texts," *Information Retrieval*, vol. 12, no. 5, pp. 526–558, 2009.

[32] A. Pak and P. Paroubek, "Twitter Based System: Using Twitter for Disambiguating Sentiment Ambiguous Adjectives," in *Proceedings of the 5th International Workshop on Semantic Evaluation*, 2010, pp. 436–439. [Online]. Available: http://www.aclweb.org/anthology/S10-1097

[33] A. Go, R. Bhayani, , and L. Huang, "Twitter sentiment classification using distant supervision," Stanford University, Tech. Rep. CS224N Project Report, 2009.

[34] A. Kiryakov, B. Popov, D. Ognyanoff, D. Manov, A. Kirilov, and M. Goranov, "Semantic annotation, indexing and retrieval," *Journal of Web Semantics*, vol. 1, no. 2, pp. 671–680, 2004.

[35] 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*, vol. 7, pp. 154–165, 2009.

[36] 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.

[37] 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*. IEEE, 2010, pp. 1–10.

[38] C. Carlos, M. Marcelo, and P. Barbara, "Information credibility on twitter," in *Proceedings of the 20th International Conference on World Wide Web*, ser. WWW '11, 2011, pp. 675–684.