EC Project 610829

A Decarbonisation Platform for Citizen Empowerment and Translating Collective Awareness into Behavioural Change

D2.3.1: Environmental Opinion Extraction

29 September 2015

Version: 1.0

Version history

Version	Date	Author	Comments
0.1	8 September 2015	Diana Maynard	Initial version
0.2	9 September 2015	Diana Maynard	improvements following MODUL's review
1.0	29 September 2015	Harith Alani	Formatting edits

Peer reviewed by: Arno Scharl, MODUL

Dissemination Level: PU – Public

This document is part of the DecarboNet research project, which receives funding from the European Union's 7th Framework Programme for research, technology development and demonstration (Grant Agreement No 610829; ICT-2013.5.5 CAPS Collective Awareness Platforms for Sustainability and Social Innovation).

Executive Summary

This deliverable provides a report to accompany the web service for environmental opinion mining delivered. The web service provides tools to perform recognition of climate-related sentiment and opinions, including the distinction between the holder of the opinion (e.g. a particular scientist) and the opinion target (what the opinion is about, e.g. fracking).

The report explains how to use the web service, describes the applications and the underlying natural language processing tools used, and details some initial experiments carried out to evaluate the performance of these tools. Finally, it provides some information about ongoing work and further possible improvements to be made.

Table of Contents

1. Introduction

This deliverable describes the ClimaPinion web service provided for environmental opinion mining in DecarboNet, and gives some more detailed information about the underlying technology of the application, along with some first experimental results to test its accuracy and effectiveness. The data annotated by these tools enables other project members to access the opinions extracted from the text, so that they can use this information within the project for experimentation; for example, the tools are used in WP4 for categorising users based on their social media behaviour (see D4.2) in order to map tweets to different stages in the 5 Doors theory of engagement.

The services require no technical skills to use, and can therefore be accessed by partners from any WP. We have also incorporated the opinion mining application not just in a web service but also in our GCP (GATE Cloud Processing) tool, which enables partners to analyse large volumes of data from the command line without having to install GATE or put pressure on the web services. Furthermore, we have enabled import and export of the documents as csv files, which means that they can be more easily integrated with other processing tools provided by the other technical partners, MODUL and OU, than the standard XML output normally produced by GATE and the GCP.

In this report, we first describe the opinion mining web service in Section 2, followed by the technology underlying the opinion mining tools in Section 3. In Section 4 we describe some preliminary evaluations we have performed on the tools and some comparison with the state-of-the-art. In Section 5 we further discuss the results of the evaluation, and outline the planned work for the second version of the software during the remainder of the project.

2. The ClimaPinion Web Service

This web service aims to annotate documents with opinions related to climate change. The web service takes as input a document or set of documents, and outputs those documents as XML files annotated with opinion information. The underlying application is developed in GATE [Cunningham 2002] and contains the following processing stages:

- standard linguistic pre-processing: tokenisation, sentence splitting, part-of-speech tagging, morphological analysis (note: these are standard GATE components which have simply been re-used without adaptation)
- environmental term extraction as provided by the ClimaTerm tools (described in D2.2 and also available as a separate web service).
- opinion extraction: identification of opinionated sentences and categorisation of their polarity, identification and categorisation of emotions, identification of opinion targets and authors where appropriate
- export as XML (inline annotation)

The opinion mining application is available via a web service at:

http://services.gate.ac.uk/decarbonet/sentiment/

The web service is publicly available; the final version will be made open source. It takes a document as input, and outputs the text as a JSON document of standoff annotations with term and URI information.

The text to be processed should be passed to the service using one of the following three request parameters.

Parameter	Supported Request	Description
text	GET or POST	Plain text to process
url	GET or POST	The URL of a document to process
file	POST	A file to process.

The response from the service is a simple JSON document containing standoff annotation markup. For example, processing the sentence "Polar bears hate climate change and so do we!" would produce the following output:


```
{
  "text":"Polar bears hate climate change and so do we!",
  "entities":{
    "SentenceSentiment":[
      {
        "indices":[
          0,
          45
        ],
        "holder":null,
        "rule2":"SentenceEntitySentiment",
        "target_string":"climate change",
        "sentiment_string":"hate",
        "emotion":"anger",
        "rule":"SentimentTerm",
        "sarcasm":"no",
        "score":-0.5,
        "polarity":"negative"
      }
    ],
    "Term":[
      {
        "indices":[
          17,
          31
        ],
        "source":"reegle",
        "rule":"Reegle",
        "label":"climate change",
        "language":"english",
        "Instance":"http://reegle.info/glossary/1018"
      }
    ]
  }
}
```


A demo is also available, where a user can type or paste in a short text and see instantly the opinion and term annotations identified. Figure 1 illustrates an example. Alongside each sentence, the main emotion identified is shown, inside a colour-coded box denoting whether it is positive (green) or negative (red). A neutral sentiment has a grey box just indicating that it is neutral.

Figure 1: Screenshot of the opinion mining demo

Note that the web service itself does not easily permit a user to input a large dataset of csv files, which is the format in which the documents are extracted from the Media Watch for Climate Change (MWCC) tools, and does not easily enable large volumes of data to be processed at once. For this purpose, we adapted our GCP (Gate Cloud Processor) standalone processing tools in order to enable this functionality. This meant that for WP4, project partners were able to run our analysis tools from the command line over their datasets. More information about this is given also in D4.2.

3. Opinion Mining Technology

3.1 Introduction and Objectives

The objective of the opinion mining technology is to perform recognition of climate-related sentiment and its classification into positive, negative and neutral polarity, as well as some core emotions such as fear, anger, joy, and so on. It includes the recognition of the holder of the opinion (e.g. a particular scientist) and the opinion target (what the opinion is about, e.g. fracking). For example, in the tweet depicted in Figure 2, we can see a negative sentiment expressed about the topic "fracking" expressed by MP Rachael Maskell. So here, the sentiment would have negative polarity, the target would be "fracking" and the opinion holder would be Rachael Maskell.

We should clarify here a point about terminology. Theoretically, opinions and sentiment are two different things, and thus by extension opinion mining and sentiment analysis. Sentiments typically express a particular polarity (positive, negative or neutral). For example "*I think your dress is pretty*" is a positive sentiment expressed by me. Opinions could express something rather more generic, e.g. "*I think that it will rain tomorrow*" is an opinion expressed by me about the weather, but it does not express any particular positive or negative sentiment. However, "opinion" can also be used to mean a positive or negative sentiment; for example, in the first example, I am expressing a positive opinion about your dress.

Figure 2: Anti-fracking tweet

In the early days of opinion mining research, the term opinion mining was thus used for something quite encompassing, while sentiment analysis was used specifically for the task of polarity detection. However, in recent years the two terms have come to be used interchangeably, in particular where sub-tasks and side-tasks have been formed (e.g. detecting whether something is opinionated or not; detecting emotions such as fear, anger etc; detecting the reliability of opinions and so on). In this deliverable, we use the term "opinion mining" to cover the tasks of detecting whether something expresses sentiment, what the polarity of the sentiment is, how strong that sentiment is, who is holding the opinion, what the opinion is about, and what emotions are being expressed. In this work, we do not attempt to distinguish opinions as a non-factual statement with neutral sentiment (as in the weather example) from a factual statement (e.g. "*it is raining*").

We also make some further clarifications about the distinction between neutral and no sentiment. Some systems make a different kind of distinction between these two cases, mainly when the system is used on longer documents. In this case, neutral is the case where there is an equal number of positive and negative elements: for example on a review site a score of 3/5 stars could be seen as equally positive and negative, where there are some good and bad points about the product. Alternatively, neutral sentiment is sometimes used to describe the case where the author clearly is expressing some sentiment but it is unclear what exactly that sentiment is. In these cases, no sentiment is different from neutral sentiment. However, it has been shown that both manual annotators and automated tools have great difficulty in distinguishing between the two cases, especially in shorter documents. In our case, looking mainly at tweets, we do not see a valid distinction between neutral and negative, and do not see a specific purpose in attempting to differentiate between the two, so we use neutral to incorporate both cases.

On the topic of **opinion holder** detection, in most cases in our scenario, the holder of the opinion is the document author, especially where the document is a tweet. This will typically be represented by a twitter username (which could reflect a person, organisation or even,

rarely, a location), but might also be the actual name of one of these, in the case of reported speech or actions (e.g. "*David Cameron says that Earth Hour is an excellent idea*" would show a positive sentiment expressed by David Cameron about Earth Hour, but would not show any particular sentiment expressed by the tweet author about Earth Hour). This level of detail is something that most opinion mining systems do not cover: they would typically associate the tweet with positive sentiment, but if we want to know the opinion of the author about Earth Hour, this would not necessarily be correct (since we do not know what the author's opinion is).

The **opinion target** is restricted to either Named Entities (Person, Location, Organisation) or to climate change terms as identified by our term extraction tool ClimaTerm (see Deliverable D2.2.1). This is because opinions about these things are considered the most relevant. However, we also classify opinionated tweets which do not have a specific target, or where the target is not one of these types, as a general opinionated statement, with no specific target. This can be useful for other purposes, such as comparing general positive vs. negative tweets, or for measuring engagement.

Most opinion mining techniques make use of machine learning (ML), but these approaches typically work best when large amounts of training data are used, for example in customer reviews where a rating system accompanies the free-form text. In particular, such approaches do not adapt well to tweets and other forms of social media [Aue 2005], especially those on a specific domain such as environmental matters. While some work in the past has focused on adapting ML methods to new domains [Balog 2006], these only really focus on the use of different keywords in similar kinds of text, e.g. product reviews about books vs. reviews about electronics. Our entity-centric approach, on the other hand, makes use of rule-based NLP techniques, but in contrast to more traditional NLP approaches involving full parsing, we use a much shallower but more focused approach based around entity and term recognition, which lends itself better to non-standard text.

3.2 General Approach

The approach used for opinion mining is a knowledge-based approach, for the reasons outlined above. Our experience has also shown that for such targeted tasks as this, a knowledge-based approach enables us more easily to make the opinion mining specific to the task: i.e. to focus exactly on the targets and opinion types, rather than just to find generic positive and negative tweets. This is also why it is hard to evaluate against other approaches, because they are not specifically adapted to the domain and task. We show in Section 4, however, some preliminary evaluations in order to situate our work in some sense against the state of the art and against human annotators. We present there a number of caveats as to why the comparison is tricky.

The application, which we call ClimaPinion, is developed in GATE, and consists of a number of components, adapted and enhanced from our generic baseline opinion mining application developed in the ARCOMEM project [Maynard 2015c]. Specifically, ClimaPinion additionally finds the relevant authors and targets, as detailed above, and breaks the opinion types down into various kinds of emotions. It also has a number of linguistic subcomponents designed to improve the analysis, namely detection of conditionals, sarcasm, swear words and so on.

Figure 3 shows a simple example of a tweet from our Earth Hour 2014 collection (see Section 4) annotated by ClimaPinion in GATE. Here the tweet shows a positive polarity, the target of the opinion is "Earth Hour" and the emotion is a happy one. In the rest of this section, we describe the various components which make up the application, and explain how they are combined.

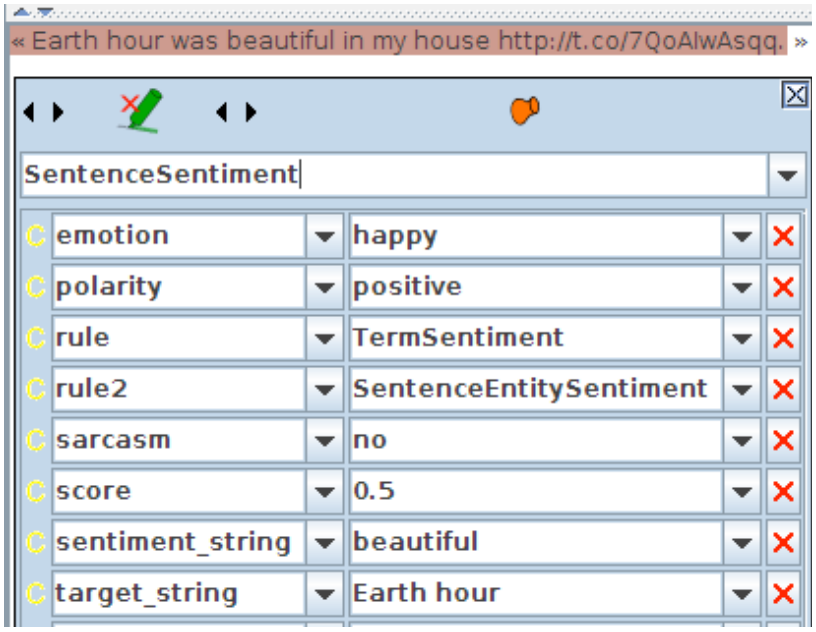


Figure 3: Example of a happy emotion annotated in GATE

3.3 Generic Opinion Mining Tools in GATE

The sentiment analysis application is designed to run on text annotated with entities and terms, and makes use of the relevant linguistic analysis associated with these. For performing generic linguistic processing and NER (Named Entity Recognition), i.e. sentences, tokens, POS tags, morphological analysis and named entities, we use the GATE ANNIE [Cunningham 2002] and TwitIE [Bontcheva 2014] tools, designed for generic text (e.g. news articles) and tweets or other social media forms respectively. The application is run conditionally on the documents, so that if the document is a tweet, it is automatically recognised as such, and TwitIE components are run on that document; otherwise, ANNIE is run on it. Following this, further linguistic processing is performed: namely term recognition and Noun Phrase and Verb Phrase chunking, which helps us identify the correct target and perform other forms of scope detection (e.g. conditional sentences, sarcasm and negation) later in the opinion mining process. These are all generic GATE tools. The reason we do not use full parsing, such as the Stanford parser, although it could potentially be helpful in giving us better relation information, e.g. in the case of scope detection, is because it tends to be very inaccurate on informal text such as tweets and other forms of social media, and also because it is incredibly slow to run. We therefore make the compromise with chunking, which tends to be more accurate and is much faster, though still not entirely error-free.

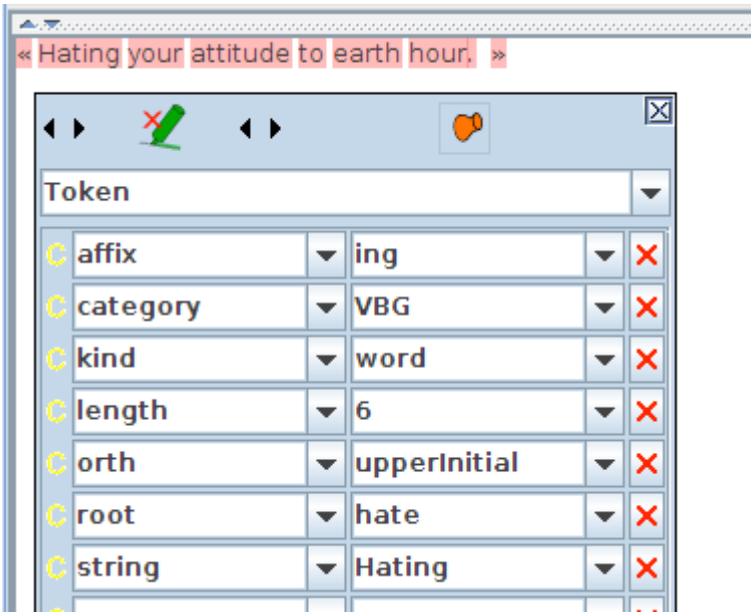


Figure 4: Example of a verb matched against its root-form

The main part of the sentiment analysis application following pre-processing comprises the following components:

- **Flexible Gazetteer Lookup**: this matches lists of sentiment words against the text. We use a flexible form of matching (the GATE extended gazetteer) which means that the words in the list are matched according to their root form. This enables different lexicalisations, e.g. plurals, different verb forms etc. to match against each other. For example, if we have the word “hating” in the text, this will be matched against “hate” in the gazetteer, because both share the same root form “hate”. Figure 4 shows an example of this (we see the word in the text “Hating” has been assigned the root form “hate”.) However, we also restrain the matching so that a match is only valid if the same part-of-speech category applies to both, i.e. a verb in the text will not be matched with an adjective from a lexicon. This is because many sentiment-bearing words have different sentiment polarity when used as different parts of speech (compare e.g. "I like it" (positive) with "someone like me" (neutral)). Surprisingly, this restriction is rarely found in other sentiment analysis tools.
- **Regular Gazetteer Lookup**: this uses a regular gazetteer, and matches lists of sentiment words against the text only if they occur in exactly the same form as the list, i.e. different lexicalisations are not matched, because these tend to be specific terms such as swear words or phrases. Multi-word phrases cannot be matched under different lexicalisations by the flexible gazetteer, so these are also matched here.
- **Sentiment Grammars:** this is a set of hand-crafted JAPE rules which annotate sentiments and link them with the relevant targets and opinion holders. They include modules for conditional sentence detection, question detection, etc.

The approach for sentiment analysis is thus a rule-based one, which is quite similar in methodology to that used by [Taboada 2011], and which is documented in [Maynard 2012]. Rather than just combining the values of any sentiment-containing words, it focuses on building up a number of linguistic subcomponents which all have an effect on the score and polarity of a sentiment. The main body of the opinion mining application involves a set of JAPE grammars which create annotations on segments of text. JAPE is a Java-based pattern

matching language used in GATE. The grammar rules use information from gazetteers combined with linguistic features (such as part-of-speech tags) and contextual information to build up a set of annotations and features, which can be modified at any time by further rules. The set of gazetteer lists contains useful clues and context words: for example, the sentiment gazetteers mentioned above have a feature denoting their part of speech, and information about the original WordNet synset to which they belong. The original lists were taken from WordNet Affect[1], but have been modified and extended manually to improve their quality: some words and lists have been deleted (since we considered them irrelevant for our purpose) while others have been added.

Once sentiment-containing words have been matched, an attempt is made to find a linguistic relation between an entity or term in the sentence or phrase, and one or more sentiment-containing words, such as a sentiment-containing adjective modifying an entity or term, or in apposition with it, or a sentiment-bearing verb whose subject or direct object is an entity. Examples could be:

- *Happy Earth Hour*: the sentiment adjective "happy" modifies the term "Earth Hour".
- *Earth Hour, a magical time:* the sentiment phrase "a magical time" is in apposition with the term "Earth Hour".
- *I love Earth Hour*: the sentiment verb "love" has the direct object "Earth Hour".

If such a relation is found, the sentence is given a Sentiment annotation, with features denoting the polarity (positive or negative) and the polarity score. The initial score allocated is based on that of the gazetteer list entry of the relevant sentiment word(s). We have seen already such an example in Figure 2.

The concept behind the scoring (and the final decision on sentiment polarity) is that the default score of a sentiment word can be altered by various contextual clues. For example, typically a negative word found in a linguistic association with a sentiment word will reverse the polarity from positive to negative and vice versa. Similarly, if sarcasm is detected in the statement, the polarity may be affected (typically, the polarity is reversed -- see the following section for more details). Negative words are detected via our Verb Phrase Chunker (e.g. "didn't") and via a list of negative terms we have compiled manually and which form another gazetteer list (e.g. "not", "never").

3.4 Decarbonet-specific modules

We have built on the generic GATE opinion mining tools described above in two main ways for the development of ClimaPinion in DecarboNet. First, we have made some specific adaptations to deal with the task and domain: this concerns the author and target detection, the addition of emotion detection, and the adaptation of sentence-level to tweet-level detection. Second, we have made some general improvements to the tools which can be used for other tasks and domains: this includes the expansion of sentiment lexicons, the addition of components such as more complex use of intensifiers, better context boundary detection, improved detection of sentiment context and so on.

Detection of opinion holders

In addition to finding the sentiment and target for each sentence, we also associate the holder of the opinion with the opinion itself. In most cases, when dealing with Twitter, the holder of the opinion is the author of the tweet. In other cases, for example if the tweet is a retweet, or if the tweet mentions an opinion held by another person, then we extract the name or username

[1] http://wndomains.fbk.eu/wnaffect.html

of that person (depending what information is available in the tweet). Figure 5 shows a screenshot in GATE of an opinion extracted from a retweet using ClimaPinion, where the holder of the opinion is the person who originally tweeted (in this case, @onsustain). Having the information about the opinion holder means that we can later perform aggregation over particular tweet authors, for example looking at all the tweets by that person which express sentiment and see how they change over time, or how their tweets are regarded by other people, what their social network (followers etc.) is like, and so on. This could lead to interesting observations.

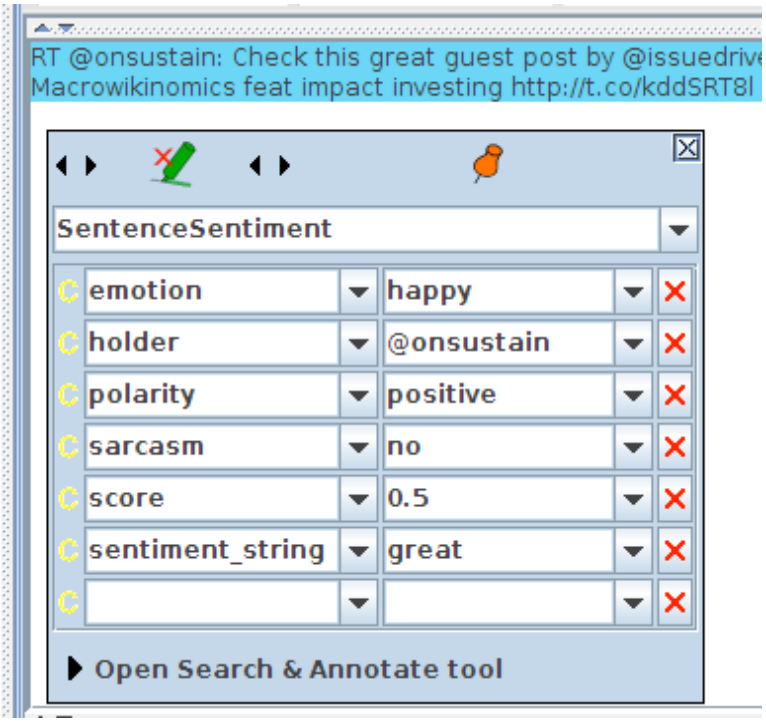


Figure 5: Annotation of an opinion and the opinion holder in ClimaPinion

Tweet-level opinion-target association

We first annotate every sentence with basic sentiment and entity/term information, as described above. All entities and terms are candidate targets: we will call these *topics*. The problem is then that we need to associate the correct sentiment with the correct topic (i.e. the target of the opinion). In the generic opinion mining application, this is performed primarily by using the closest topic within the same phrase or sentence chunk. Furthermore, in that application, only topics within the same sentence as the sentiment are matched, and if there are multiple topics or sentiments, the nearest combination is matched and the others are ignored.

In the DecarboNet application, we do things a little differently. We collect the following information: number of sentiments, topics and the position of the sentiment in the tweet. Then we apply a context algorithm:

- If the tweet contains one or more Topics and one or more Sentiments with the same polarity (positive or negative), then a SentenceSentiment annotation is created. If there are multiple sentiments, we use the one with the highest score, or failing that, the nearest one to the topic.
- If the tweet contains one Topic and more than one Sentiment with different polarity, then we take into account the Sentiment from the same sentence as the Topic.

- If the tweet contains more then one Topic and more than one Sentiment with different polarity in the same sentence, then we build what we call a topic context.
- We build topic contexts for each topic as follows: for the first topic encountered, the context is from the beginning of the sentence until the second topic. The rest of the Sentence will be the context for the second (or third etc.) topic.
- We also use some phrase breaker words like "but", "because" etc. in order to delimit a phrase which should end the context, i.e. a context cannot span two phrases joined by such words.

Note, however, that in some cases, we override this topic selection method by preferring some topics over others. In the case of the Earth Hour tweets, we tend to prefer the term "Earth Hour" as the target of the opinion, since the opinions almost always refer to this even when it is not necessarily the closest topic to the sentiment words. An example of this would be where people talk about Earth Hour in their particular location, or when they talk about saving energy (a term which would normally be a topic for consideration as a target) during Earth Hour. In these cases we prefer to use Earth Hour as the opinion target.

Figure 3 earlier showed an example of a tweet containing an opinion and a target ("*Earth Hour was beautiful in my house"*), where the opinion was positive about the target *Earth Hour*. Figure 6 below shows another example -- here the tweet contains a negative opinion about the target *climate change*.

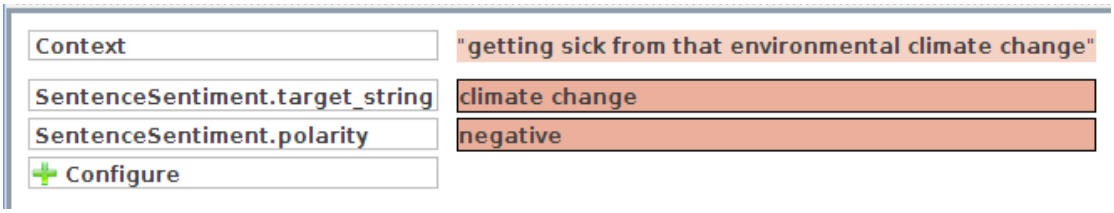


Figure 6: Example of an opinion and target in GATE

Sentiment Aggregation

Sentiment aggregation is carried out by groovy scripts which combine the scores for sentiments over sentences (and potentially paragraphs and documents), and output an aggregated score for each. This is essentially the average score for the document, and is a standard way to perform this task. Note that in the literature, improvements to document-level opinion finding have been achieved by a method which traverses the document one sentiment at a time and alters the score sequentially using a directed transition graph [Rocha 2015]. This is pointless, however, on short documents such as tweets. In fact, trying to provide an average opinion for a long document is largely not used these days[2], since it does not prove very useful for insight or further analysis. Other ways of interpreting the data are likely to give better insight, for example measuring the opinionatedness of a document and the range of opinions expressed might be useful in analysing debated topics or comments on a news article [Maynard 2014b].

Expansion of Sentiment Lexicons

One of the problems we found with our existing sentiment lexicons was that they were quite incomplete, and this accounted for low Recall. We therefore developed a methodology to

[2] http://www.kdnuggets.com/2015/08/11-things-about-sentiment-analysis.html

expand the lexicon with domain-appropriate additions. Over a corpus of domain-specific tweets, we check every adjective/adverb/noun/verb for sentiment, and for each one, we extract synonyms from WordNet. The synonyms are checked for sentiment in our lists, and any new ones added to the lexicon, with the same score as the related words. If no synonyms are found, first order hyponyms are investigated in the same way.

We tested the expansion tool using the Earth Hour 2014 corpus. The tool generated 351 new sentiment terms, although this includes morphosyntactic variants (for example, "embrace", "embraces" and "embraced" were all generated separately). We will investigate in future how to adapt the tool to consider only the root form, and we will perform more experiments with larger corpora to better see the effect of lexicon expansion.

Intensifier Detection

This involves modifying the score of the sentiment words to increase their strength. Typically adverbs preceding a sentiment adjective will increase the strength, e.g. "very boring" is stronger than "boring". Sometimes, however, they will decrease the strength, e.g. "quite boring" is slightly weaker than "boring". Some adjectives which themselves do not typically carry sentiment, can also act as intensifiers over sentiment-containing nouns, e.g. "utter rubbish" is stronger than "rubbish". Finally, swear words typically convey negative sentiment when they occur without the presence of a sentiment-bearing word. However, when they occur in combination with a sentiment-containing adjective, they act as intensifiers. For example, one can strengthen both "amazing" and "awful" by preceding it with one's chosen swear word. We have thus incorporated some sets of intensifiers into the algorithm, according to these different categories.

Emotion Detection

Finally, we have also classified the sentiments into different emotion types. It is important to note that these are not necessarily classic emotion types[3], but rather types that we have deemed most useful for our analysis in the project case studies, in particular with respect to the identification of user engagement and the 5 doors theory of change described in WP4. Therefore we include not only standard emotions such as joy, fear and anger but also types such as "cute" and "swearing", which are useful for this purpose.

Every sentiment therefore also has a feature called "emotion". Negative sentiments are categorised as one of:

- anger
- disgust
- fear
- sadness
- bad (a generic negative category for anything not captured by the previous negative emotions)
- swearing (note that swearing can also be positive when used as an intensifier as explained previously. In this case it is not listed here but falls under one of the positive emotions).

[3] See e.g. Aristotle's list of emotions http://spot.colorado.edu/~hauserg/ArEmotList.htm or Robert Plutchik's theory of emotions [Plutchik 2011].

Positive emotions are classified as one of:

- joy
- surprise
- cheeky
- happy
- cute
- good (as with "bad", a generic category that captures all other positive emotions not otherwise classified)

The emotion classification is performed primarily by means of the gazetteers. The sentiment lexicons used were derived from the NRC Emotion Lexicon (EmoLex) [Mohammad 2013], which was created via crowdsourcing and which categorises words into one of 8 types: anger, anticipation, disgust, fear, joy, sadness, surprise and trust. The anticipation and trust categories were not used as they do not directly translate into positive and negative polarities. We added the extra categories happy, cheeky, cute, and swearing, as these were deemed useful for the purposes of the project. The emotions are primarily used in the project in order to categorise tweets according to the type of engagement users have with environmental topics (performed in WP4), which is why they differ slightly from the original categories used in EmoLex and by other traditional categorisation schemes. Further work is needed to analyse these and to add/amend as necessary once we are aware of their utility in the engagment categorisation. Emotion words belonging to the extra categories were added manually after inspection of some large corpora, as were some extra terms to the existing EmoLex categories (we added some manually and some from other existing lexicons such as WordNetAffect [Strappavara 2004]. We have not yet run any experiments to analyse the correctness of the categorisation, but this is planned for the second phase. Figure 7 shows an example of some sample sentences classified with emotions, as performed in the web service demo.

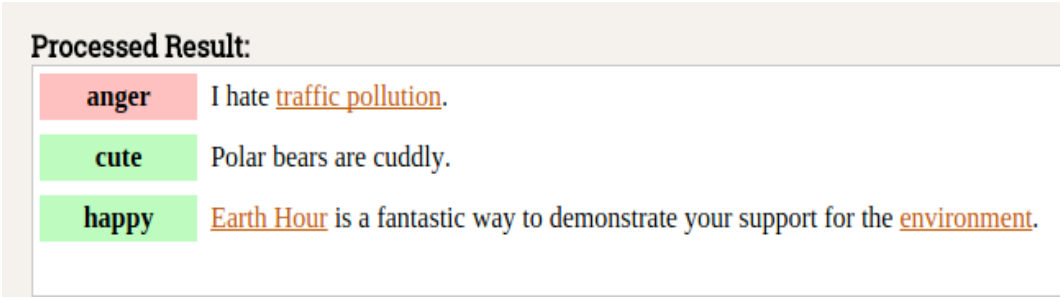


Figure 7: Examples of emotion classification

3.5 Scientific Novelty

As described above, the development of the ClimaPinion opinion mining tools for this task largely builds on our existing GATE framework for text analysis, and improves on it in a number of ways. In this sense, we have advanced the state of the art in opinion mining by combining a number of linguistic features and by adapting the sentiment analysis specifically to the domain of environmentally-related social media and the task at hand. For example, the target of the opinion should be an entity or a term relevant to climate change. One of the biggest differences from typical opinion mining tools is that we do not try to find just the overall sentiment of the tweet, but the much more specific target-related approach. We also specifically relate the opinion holder of the opinion and target.

There are a number of specific components which exhibit scientific novelty: the lexicon expansion techniques, the addition of linguistic components for detecting sarcasm, specific use of swear words, conditionals and so on -- in particular, not just recognising that these things exist (e.g. that sarcasm is present) but also in determining how this affects the sentiment expressed (i.e. by finding the scope and by correctly understanding how sarcasm, conditional sentences etc. change the polarity or extent of the sentiment). Note that the sarcasm work was commenced by us in a previous EU project ARCOMEM, and is documented in [Maynard 2014a] but has been enhanced and improved in DecarboNet. The evaluations described in Section 4 demonstrate further some improvements on current state-of-the-art systems, and we expect to see greater improvements in the second phase of the work.

On a wider level, our framework for social media analysis goes beyond the state of the art by combining many different components into a single system, in a flexible and easily extendable architecture. This is demonstrated clearly in [Maynard 2015a]. Unlike most existing opinion mining tools, the methodology and results are completely transparent -- this means that when errors occur, the system can be tweaked and improvements can be made easily. With machine learning based tools, it is often hard to make improvements except by random trial and error with additional features. We have also experimented successfully with adapting the framework and components to a slightly different task (political tweets) described in [Maynard 2015b] and [Dietzel 2014].

4. Evaluation

While our tools are designed to be a little different from generic opinion mining tools, we nevertheless need to evaluate them against some benchmarks, because absolute accuracy figures are not entirely meaningful on their own, especially since opinion mining is such a hard and varied task. We therefore compared the tools not only against some gold standard data, but also against three other systems: two pre-DecarboNet baselines (ARCOMEM and DIVINE) and the SentiStrength tool, described below. These three systems have been designed for generic opinion mining tasks and have not been specifically adapted to the domain, although they have all been previously tested and evaluated on tweets. ClimaPinion in contrast uses more sophisticated linguistic technology, dealing with issues such as conditional sentences, negation scope, sarcasm, questions and so on, which can have considerable impact on the way sentiment-containing words should be interpreted [Maynard 2014]. The evaluation thus investigates to what extent these kind of additions are useful.

The first pre-DecarboNet baseline **ARCOMEM** [Maynard 2015c, Maynard 2012] is an opinion mining tool that was developed in GATE for use in the EU ARCOMEM project[4]. It essentially comprises the core GATE opinion mining tools before the enhancements described in this deliverable were developed. This acts as a good baseline for the GATE development; it is not tuned to the environmental domain and is less sophisticated, but uses the same essential principles as the ClimaPinion tool.

The second pre-DecarboNet baseline we use [Gindl 2010] was developed in the **DIVINE** project[5], and is based on the aggregation of the sentiment scores of any sentiment-containing words in the sentence or document, using a large lexicon of sentiment words and their scores. The lexicon is compiled from the tagged dictionary of the General Inquirer, containing 4,400 positive and negative sentiment words [Stone 1966], and extended by adding linguistic variants of these terms, such that the complete lexicon contains around 7,000 terms with semantic orientation. The lexicon is thus much larger than that used by the ClimaPinion

4 http://www.arcomem.eu

5 https://www.weblyzard.com/divine/

tool, but in contrast, less linguistic analysis is done on the text itself and more reliance is made on the lexicon.

SentiStrength [Thelwall 2010] is a freely available tool for opinion mining used by a number of researchers as well as in some business applications. It is designed to estimate the strength of positive and negative sentiment in short texts, and deals well with informal language such as tweets. It is claimed to have human-level accuracy [Thelwall 2012] on such texts (except for political texts). Unlike most other tools, SentiStrength reports two sentiment strengths separately: negativity on a scale of -1 to -5 (where -5 is extremely negative), and positivity on a scale of 1 to 5 (where 5 is extremely positive).

To make the evaluation procedure as easy as possible, we developed a GATE plugin for the Java version of SentiStrength, which we have made publicly available via the SentiStrength website[6]. The plugin is customisable according to the various parameters, but in the default setting used in our experiments, the total positive, negative and combined score is output for each Sentence in the document. The combined score is simply the sum of the positive and negative scores, e.g. a positive score of +2 and a negative score of -1 would have a combined score of +1. For our experiments, we further added a text-based feature whose value can be "negative", "positive" or "neutral" in order to correlate better with our own system output, since it would have been difficult to get a meaningful comparison between the actual numerical scores of our system and SentiStrength's. Furthermore, the numerical score of our ClimaPinion system is far from fully-fledged and acts currently only as a rough indicator of the strength of opinion; this is something that we plan to develop more fully in the second version.

Note also that our experiments assess only the detection of polarity (positive, negative and neutral) but not the association between sentiment and the opinion holder and targets, since the other tools do not have this functionality. We leave this for future work in the next phase, along with the evaluation of the emotion detection.

Corpus 1: SentiStrength Twitter corpus

Our first experiment compares ClimaPinion with SentiStrength and ARCOMEM on a corpus of 4242 tweets made available by Mike Thelwall[7] and on which SentiStrength has previously been evaluated [Thelwall 2012]. Note that the scores we obtained with SentiStrength on this corpus may differ from those previously reported or obtained by others, due to the settings we used for it (default settings) and due to the way in which we did the comparison, as detailed above. Table 1 shows the results obtained. We can see that while ClimaPinion performs significantly better than ARCOMEM, it actually does not perform quite as well as SentiStrength on this corpus. However, there are a number of reasons for this.

Tool	Correct	Incorrect	Accuracy
SentiStrength	2510	1732	59.17 %

6 http://sentistrength.wlv.ac.uk/

7 Downloaded from http://sentistrength.wlv.ac.uk/

ClimaPinion	2427	1815	57.21 %
ARCOMEM	1953	2289	46.04 %

Table 1: Comparison of ClimaPinion and SentiStrength on Corpus 1

The first thing to note is the way the corpus was annotated, and the assumptions made by SentiStrength. Without contrary evidence, posting a URL is annotated as a positive tweet in the gold standard, since it is claimed that people generally post URLs in order to endorse them. This is not, however, necessarily the case, since people also sometimes post URLs for general discussion or even to show outrage, and we do not assume any sentiment unless more explicitly demonstrated in the text. This accounts for a high proportion of the mismatch between SentiStrength's and our tool's performance. Other instances where we disagree with the gold standard annotations are constructions such as conditionals which demonstrate irrealis mood. For example, in the gold standard, the tweet "I'd like to be in the midst of it all" is marked as positive, but we do not feel this is a positive tweet (since the author would be happy if they were in the midst of it, but they are not). Similarly, tweets such as "I need a nice tea-drinking pic" are annotated as positive in the gold standard, but we feel this is equally wrong. Finally, we should note that this corpus is a general twitter corpus, and is not specifically about the environmental domain, to which our ClimaPinion tool is tuned.

	ClimaPinion Negative	ClimaPinion Neutral	ClimaPinion Positive
Key Negative	304	532	113
Key Neutral	154	1458	341
Key Positive	80	595	665

Table 2: Confusion matrix for ClimaPinion on Corpus 1

	SS Negative	SS Neutral	SS Positive
Key Negative	449	326	174
Key Neutral	257	1038	658
Key Positive	93	224	1023

Table 3: Confusion matrix for SentiStrength on Corpus 1

If we look at the confusion matrices shown in Tables 2 and 3, we also see an interesting distinction. SentiStrength classifies far fewer tweets than ClimaPinion as neutral, so in terms of finding which tweets are opinionated, it scores high on Recall but low on Precision overall (i.e. it overclassifies many tweets as opinionated). ClimaPinion, on the other hand, is very conservative about classifying tweets as opinionated, because it is designed to only classify them if the confidence level is quite high. So ClimaPinion scores low on Recall but high on Precision overall. In the same way, SentiStrength also misclassifies many positive tweets as negative and vice versa, while ClimaPinion misclassifies far fewer tweets in this way. In summary, SentiStrength has greater accuracy on positive and negative tweets than

ClimaPinion, but worse accuracy on neutral tweets, i.e. it tries to assign sentiment where there is none.

Corpus 2: Earth Hour 2014

For the second experiment, we manually annotated a corpus of 500 tweets about Earth Hour 2014, which were randomly selected from a larger set that had been previously collected in WP4. We then compared the ClimaPinion, ARCOMEM, DIVINE and SentiStrength tools against this gold standard set. Results are shown in Table 4.

It is immediately evident that results on this dataset are much higher for all systems than on the general corpus used in the first experiment. There are several reasons for this. First, we believe that our gold standard annotations are more realistic: as mentioned above, we do not, for example, annotate a simple pointer to a URL as a positive instance because one cannot really be sure about this even if most references to URLs in tweets are positive. So we annotate a tweet as sentiment-containing only if it is clear that this is really true. Second, the tweets are domain-specific in this experiment, and are thus more focused, which means that one can make better predictions and also that there is less ambiguity within the corpus (though there is still just as much ambiguity between the use of words in the corpus and the use of words in general: for example, if we talk about energy we are very likely to be talking about the environmental sense, but there is still ambiguity between this and other senses of energy, which may impact the use of lexicons and so on). Third, we note that while the results for all systems are higher than for the first experiment, there is also a more noticeable difference between the performance of SentiStrength and ClimaPinion. This might be because the ClimaPinion system has been developed specifically for this domain (in particular, with the kinds of sentiment words that are used in talking about things like Earth Hour). This reflects also the large discrepancy between ARCOMEM and ClimaPinion.

Tool	Correct	Incorrect	Accuracy
ClimaPinion	434	66	86.80 %
DIVINE	257	101	79.80 %
ARCOMEM	351	147	70.34 %
SentiStrength	331	169	66.20 %

Table 4: Evaluation on Earth Hour 2014 corpus

Table 5 shows a confusion matrix for ClimaPinion on this corpus. We can see here that, unlike with the SentiStrength general twitter corpus, here the biggest source of confusion for our system was in falsely detecting positive and negative opinions which should have been neutral. As before, there was very little confusion between positive and negative, in either direction; most of the confusion was between positive/negative and neutral, i.e. opinionated or not.

	ClimaPinion Negative	ClimaPinion Neutral	ClimaPinion Positive
Key Negative	32	11	13
Key Neutral	32	275	44
Key Positive	3	9	79

Table 5: Confusion Matrix for ClimaPinion on Corpus 2

To investigate this further, we also performed an **error analysis** in order to understand what kind of errors the ClimaPinion system was making. We classified the errors into 9 types as follows:

- *missing sentiment-containing words or expressions*: for example, the word "involuntary" when associated with Earth Hour usually has a negative connotation, but it was missing from our lexicon
- *spelling errors in sentiment words*: for example, the word "celeberate" was used instead of "celebrate", and was not therefore not found in the lexicon (note: ClimaPinion does perform some normalisation of slang, but this is more for single duplicated letters and common slang terms rather than just poor spelling).
- *made-up sentiment words*: for example, the word "Awoooh" is recognisable to a human as denoting positive sentiment, but would not be found in any lexicon as it is both invented and also contains many duplicated vowels (numerous variants of the vowel duplication would be possible).
- *sentiment words used out of context*: for example, the word "best" was listed in our lexicon as a positive word, but it was used also in the phrase "you'd best be back soon" where it did not denote a positive sentiment.
- *incorrect hashtag decomposition*: for example, the hashtag #uselesspayless was wrongly tokenised by TwitIE as "useless pay less" instead of "use less pay less", and the word "useless" was annotated as negative.
- *missing sarcasm detection*: for example, the tweet "And we are officially celebrating Earth Hour after each hour. #PMSL sucks. Again." was annotated as positive because we incorrectly associated "celebrate" with "Earth Hour", whereas we should have combined also the negative aspect of "sucks" and detected that the first sentence was sarcastic, giving us a negative spin on the situation.
- *incorrect linguistic pre-processing*: for example, a verb classified as a noun is not matched with the sentiment word in the lexicon by our tools if it is listed as sentiment-containing only when a verb.
- *impossible to ascertain the sentiment automatically*: some cases were too obscure for a system to be able to detect without a lot of world knowledge. For example "Turning off power! Earth Hour!" was deemed to be a positive tweet about Earth Hour, but there were no reliable clues in the text itself.
- *non-English tweets* where our system nevertheless attempted to classify the opinion

Figure 8 shows the percentage frequency of the different kinds of error. We can see that the most frequent errors were caused by missing lexicon entries and sentiment words used out of context. A substantial number were also caused by the use of foreign language (especially in the case where the tweet was mixed language) and where the sentiment was virtually impossible to obtain automatically.

Most of these errors, other than the missing lexicon entries, are actually quite hard to resolve, especially if we want to maintain our high Precision performance. The most obvious error to fix is the lack of lexicon entries, which is something we will investigate further in the rest of the project.

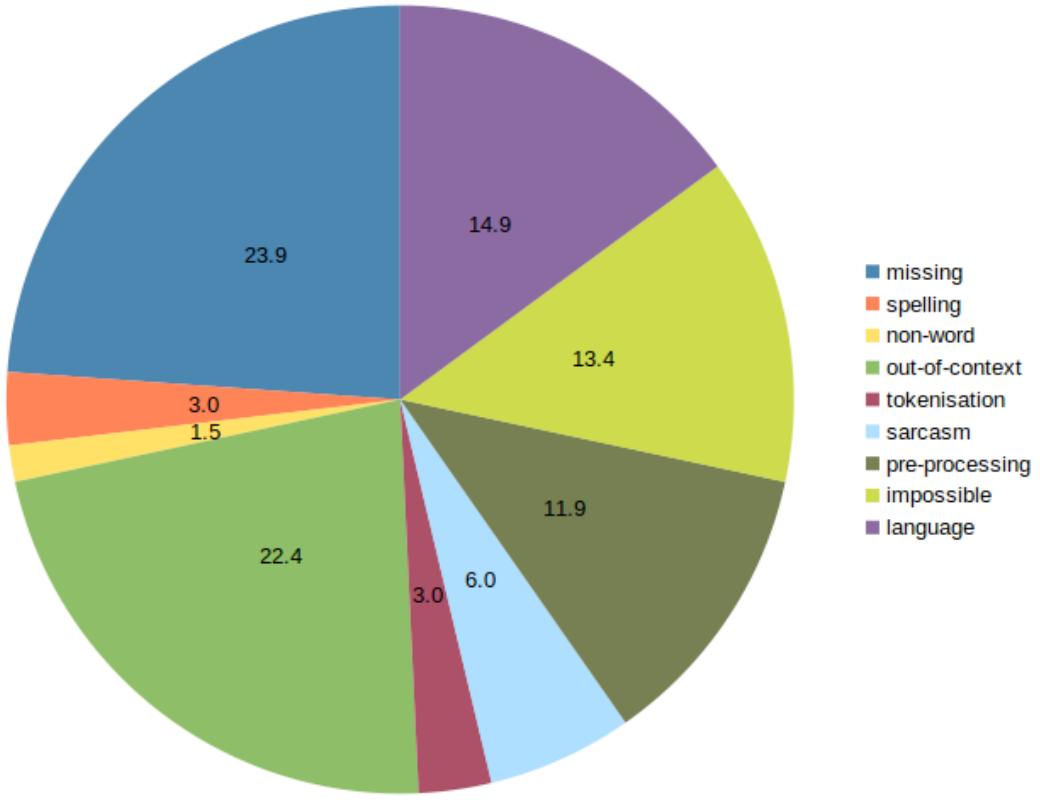


Figure 8: Error types by percentage

Corpus 3: Earth Hour 2015

Since Corpus 2 (Earth Hour 2014) was developed by only one annotator and therefore could be biased, we created a more objectively annotated corpus using crowdsourcing and triple annotation. For this we selected at random 600 tweets from the Earth Hour 2015 dataset (collected by WP4), removing any non-English tweets that had accidentally crept into the collection. Using GATE's crowdsourcing plugin [Bontcheva 2014] we assigned the dataset to a number of annotators, such that each tweet was triple-annotated. In total, there were 16 annotators, who annotated between 50-200 tweets each. Each annotator was limited to a maximum of 200 tweets, so that the set would not be too biased by a single annotator and so that annotators would not become bored and therefore make mistakes. The annotators were not all native English speakers, but were all fluent in English and had a good understanding both of the task and of the climate change domain and Earth Hour.

The crowdsourcing plugin also enables consensus making after the annotation phase is complete, using a majority vote system. Since there were 3 possibilities for any tweet (positive, negative or neutral), in the case of a 3-way tie, the decision was made by an independent arbitrator. This was the case for only 4 tweets out of 600, and were all quite easily resolvable. Appendices E and F show the task and instructions given to the annotators.

Table 6 shows the results for the 4 systems. Here we see ClimaPinion score the highest, closely followed by SentiStrength. Interestingly, this differs from the second evaluation on the Earth Hour 2014 dataset, where SentiStrength performed much worse comparatively, though with roughly the same actual accuracy score (around 65%). For some reason, the other 3 systems all perform worse on this dataset than on the Earth Hour 2014 one. We investigated the results and the annotations a little more closely.

Tool	Correct	Incorrect	Accuracy
ClimaPinion	398	202	66.33 %
SentiStrength	390	210	65.00 %
DIVINE	360	240	60.00 %
ARCOMEM	287	313	47.83 %

Table 6: Evaluation on Earth Hour 2015 corpus

It was sometimes not clear even to the human annotators what the tweet meant or what kind of message was being portrayed. For example, with the tweet:

"To celebrate the end of Earth Hour 2015, I simulated a Federal Signal 3T22A sounding off in alternating wail."

one annotator commented that they did not understand it and were not sure if it was sarcastic, while the other two annotators deemed it neutral (probably because the instructions told them to annotate anything with no clear sentiment as neutral).

Inter-annotator agreement was measured using Fleiss' kappa, and produced a score of 44.19. Note that there is no generally agreed measure of significance for this; according to [Landis 1977] our score indicates moderate agreement, though this is by no means universally accepted. It should be pointed out also that the number of categories affects this score. We use Fleiss' kappa rather than the more traditionally used Cohen's kappa for intern-annotator agreement, because the latter can only be used between two raters, whereas we have three raters. While the kappa score is quite low, recall that we use the majority judgement on the tweets, so the fact that one out of three annotators did not agree is not so important, other than to emphasise the difficulty of the task.

The proportion of judgements is interesting: positive and neutral were much more frequent than negative, as shown in Figure 8. We found this to be typically the case with tweets about Earth Hour, because people posting about it are either simply informing, or are sharing positively. The people who do not care about Earth Hour typically do not bother tweeting about it.

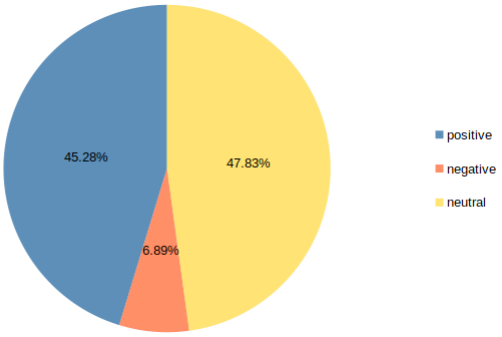


Figure 9: Distribution of polarity in corpus 3

The co-occurrence matrix in Table 7 shows how often annotators agree for each of the three polarity types, and which types they confuse. Note that this matrix should be interpreted a little differently from the confusion matrices in Tables 2, 3 and 5, which show how two different sets of answers compare against each other. In this matrix, because we are comparing three sets of answers, we show the co-occurrence for each polarity type for each

tweet (so one cannot read either the rows or columns as being the "correct" answer as one can with the previous confusion matrices, and one could swap the sets of rows with the sets of columns and get the same result).

	Negative	Neutral	Positive
Negative	62	35	9
Neutral	35	426	244
Positive	9	244	396

Table 7: Co-occurrence matrix for human annotators

Table 8 shows the confusion matrix for our system compared with the gold standard provided by the annotators. We can see clearly that the biggest source of confusion (just over 46% of errors) was where the correct answer was positive but our system found no sentiment. The second biggest source of confusion was where the correct answer was neutral but our system found a positive sentiment (33%). In total, this means 70% of errors were caused by neutral/positive confusion, correlating well with the human judgement problems where 88% of errors were caused by neutral/positive confusion. In contrast, less than 10% of errors in our system were caused by negative/positive confusion (in either direction), and only 11% were caused by negative/neutral confusion (in either direction). This all bodes well for future improvements to the system, which will include further discussion with project partners and better clarification of guidelines and the positive/neutral distinction.

	ClimaPinion Negative	ClimaPinion Neutral	ClimaPinion Positive
Key Negative	12	19	9
Key Neutral	4	217	66
Key Positive	10	94	169

Table 8: Confusion matrix for ClimaPinion compared with human annotators

We can also calculate from Table 8 the individual Precision and Recall for each value of the polarity. This is shown in Table 9. Here what we can see is that our best recall is for neutral, and our best precision is for positive polarity. Our worst precision and recall is for negative polarity; however, from the pie chart in Figure 9, we know that the frequency of negative tweets in the corpus is extremely low compared with that of positive and neutral tweets.

ClimaPinion Polarity Value	Precision	Recall
Negative	46.15	30.00
Neutral	65.76	75.61
Positive	69.26	61.91

Table 9: Precision and Recall for polarity values

We see also from Table 8 that there was very little confusion between negative and positive, and not much confusion between negative and neutral, but great confusion between positive

and neutral. This is easy to understand, because many tweets were not overtly positive but nevertheless could be understood to endorse Earth Hour in some way (for example, generally talking about Earth Hour can be seen as promoting the campaign if nothing explicitly negative is mentioned). In our system, we do not try to annotate such things as positive, but some of the annotators seemed to find this a difficult distinction to make. In some sense, the distinction between negative and positive, and between negative and neutral, is the most important to be clear about; if we look at discussions about engagement with the concept of climate change and the topic of the environment, as shown in WP4, it is clear that the distinction between an overtly positive tweet and a neutral tweet about the topic is actually not so important. In some sense, therefore, absolute figures for accuracy are less important than considering the confusion matrix for the system and how well it performs on correctly separating negative tweets from neutral and positive ones.

5. Summary and further work

In this deliverable we have described the first version of our tools for opinion mining which reveal the sentiments expressed by the public about climate change-related issues. The tools have been made available for use within the project both as a web service and via our GCP tool which enables large-scale processing in a format easily accessible to users (csv input and output).

In addition to applying the opinion mining tools to data from Earth Hour in WP4 for the purposes of identifying user engagement issues, we have also applied the tools to a political dataset outside this project, in order to study the engagement of citizens with respect to tweets by politicians and election candidates leading up to the UK elections [Dietzel 2015] [Maynard 2015]. Results showed that climate change and environmental topics typically engage UK citizens more than most other political topics. In all likelihood, this is because people feel that climate change is a topic about which they can pro-actively help to mitigate the adverse effects, unlike topics such as immigration and the economy. Both sentiment analysis and term detection (as described in the previous deliverable D2.2.1) played a key role in this engagement study, since it has been shown that factors such as opinionated tweets, and particularly positive tweets, are an important indicator of user engagement, as described in D6.2.1 and in [Rowe 2014].

The initial results from the Earth Hour evaluations are promising. While on the crowdsourced corpus, our tool does not achieve much higher accuracy overall than SentiStrength, one of the most widely used state-of-the-art tools for sentiment detection, it does achieve better precision overall, which is by design rather than accident. Our tools have been developed specifically to only predict a sentiment where confidence is high. We consider this to be a more useful approach when trying to draw conclusions about the classification of users and their attitudes towards climate-change related topics such as Earth Hour. Clearly, there are many improvements still to be made to the tools, and this will be the focus of the remaining work on this task, along with further evaluations (including assessing the accuracy of the opinion holder and opinion target association, and emotion detection). The final version of the tools will be fully integrated into the project architecture rather than remaining as stand-alone services.

A. List of Figures

B. List of Tables

C. List of Abbreviations

Abbreviation	Explanation
CA	Consortium agreement
DoW	Decription of work, i.e. GA - Annex I
EC	European commission
GA	Grant agreement
IP	Intellectual property
IPR	Intellectual property rights
PC	Project coordinator
PMB	Project management board
SC	Scientific Coordinator
PO	Project officer
PSB	Project steering board
DM	Data Manager
AB	Advisory board
WP	Work package

D. References

[Aue 2005] A. Aue and M. Gamon. 2005. Customizing sentiment classifiers to new domains: a case study. In Proc. of the International Conference on Recent Advances in Natural Language Processing, Borovetz, Bulgaria.

[Balog 2006] K. Balog, G. Mishne, and M. de Rijke, "Why are they excited?: identifying and explaining spikes in blog mood levels," in *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, Stroudsburg, PA, USA, 2006, pp. 207–210.

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

[Bontcheva 2014] K. Bontcheva, I. Roberts, L. Derczynski, D. Rout. The GATE Crowdsourcing Plugin: Crowdsourcing Annotated Corpora Made Easy. Proceedings of the meeting of the European chapter of the Association for Computation Linguistics (EACL).

[Cunningham 2002] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.

[Cunningham 2011] H. Cunningham, V. Tablan, I. Roberts, M. A. Greenwood, & N. Aswani (2011). Information extraction and semantic annotation for multi-paradigm information management. In Current Challenges in Patent Information Retrieval(pp. 307-327). Springer Berlin Heidelberg.

[Dietzerl 2014] A. Dietzel and D. Maynard. Climate Change: A Chance for Political Re-Engagement? In Proc. of the Political Studies Association 65th Annual International Conference, April 2015, Sheffield, UK.

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

[Maynard 2014a] Diana Maynard and Mark A. Greenwood. Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. Proc. of LREC 2014, Reykjavik, Iceland, May 2014.

[Maynard 2014b] Diana Maynard, Gerhard Gossen, Marco Fisichella, Adam Funk. Should I care about your opinion? Detection of opinion interestingness and dynamics in social media. Journal of Future Internet, Special Issue on Archiving Community Memories, 2014.

[Maynard 2015a] D. Maynard and K. Bontcheva. Understanding climate change tweets: an open source toolkit for social media analysis. In Proc. of EnviroInfo 2015, Copenhagen, Sep. 2015.

[Maynard 2015b] D. Maynard, M. A. Greenwood, I. Roberts, G. Windsor, K. Bontcheva. Real-time Social Media Analytics through Semantic Annotation and Linked Open Data. Proceedings of WebSci 2015, Oxford, UK

[Maynard 2015c] Diana Maynard and Jonathon Hare. Entity-based Opinion Mining from Text and Multimedia. In "Advances in Social Media Analysis", Mohamed Gaber, Nirmalie Wiratunga, Ayse Goker, and Mihaela Cocea (eds.) 2015, Springer.

[Mohammad 2013] Crowdsourcing a Word-Emotion Association Lexicon, Saif Mohammad and Peter Turney, *Computational Intelligence,* 29 (3), 436-465, 2013.

[Plutchik 2001] Plutchik, Robert. "The Nature of Emotions Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice." American Scientist, 89.4 (2001): 344-350.

[Rocha 2015] Rocha, Leonardo, Fernando Mourão, Thiago Silveira, Rodrigo Chaves, Giovanni Sá, Felipe Teixeira, Ramon Vieira, and Renato Ferreira. "SACI: Sentiment analysis by collective inspection on social media content."Web Semantics: Science, Services and Agents on the World Wide Web (2015).

[Rowe 2014] Rowe, M., and Alani, H., 'Mining and Comparing Engagement Dynamics Across Multiple Social Media Platforms' in Proceedings of the 2014 ACM Conference on Web Science, (2014), pp. 229 – 238

[Scharl 2013] Arno Scharl, Alexander Hubmann-Haidvogel, Albert Weichselbraun, Heinz-Peter Lang, Marta Sabou (2013). Media Watch on Climate Change -- Visual Analytics for Aggregating and Managing Environmental Knowledge from Online Sources, 46th Hawaii International Conference on System Sciences, pp. 955-964.

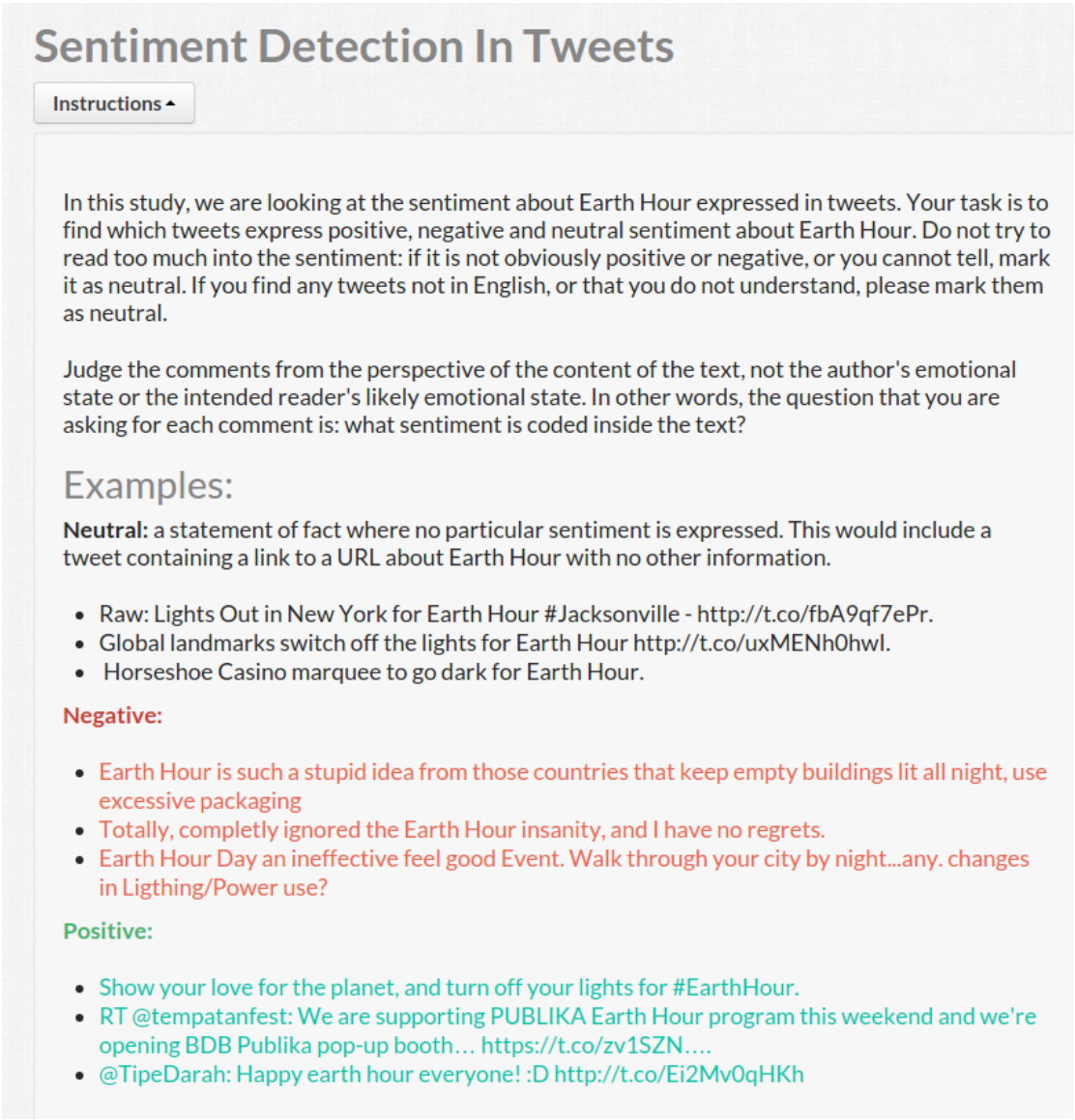
[Scharl 2014] Scharl, A., Kamolov, R., Fischl, D., Rafelsberger, W. and Jones, A. (2014). Visualizing Contextual Information in Aggregated Web Content Repositories. 9th Latin American Web Congress (LA-WEB 2014). Ouro Preto, Brazil: Forthcoming.

[Strappavara 2004] C. Strapparava and A. Valitutti. WordNet-Affect: an affective extension of WordNet. In Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, 2004.

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

[Weichselbraun et al. 2014] Weichselbraun, A., Streiff, D. and Scharl, A. (2014). Linked Enterprise Data for Fine Grained Named Entity Linking and Web Intelligence. 4th International Conference on Web Intelligence, Mining and Semantics (WIMS-2014). Thessaloniki, Greece: ACM Press.

E. Instructions for Annotators



Sentiment Detection In Tweets

Instructions

In this study, we are looking at the sentiment about Earth Hour expressed in tweets. Your task is to find which tweets express positive, negative and neutral sentiment about Earth Hour. Do not try to read too much into the sentiment: if it is not obviously positive or negative, or you cannot tell, mark it as neutral. If you find any tweets not in English, or that you do not understand, please mark them as neutral.

Judge the comments from the perspective of the content of the text, not the author's emotional state or the intended reader's likely emotional state. In other words, the question that you are asking for each comment is: what sentiment is coded inside the text?

Examples:

Neutral: a statement of fact where no particular sentiment is expressed. This would include a tweet containing a link to a URL about Earth Hour with no other information.

- Raw: Lights Out in New York for Earth Hour #Jacksonville - http://t.co/fbA9qf7ePr.
- Global landmarks switch off the lights for Earth Hour http://t.co/uxMENh0hwl.
- Horseshoe Casino marquee to go dark for Earth Hour.

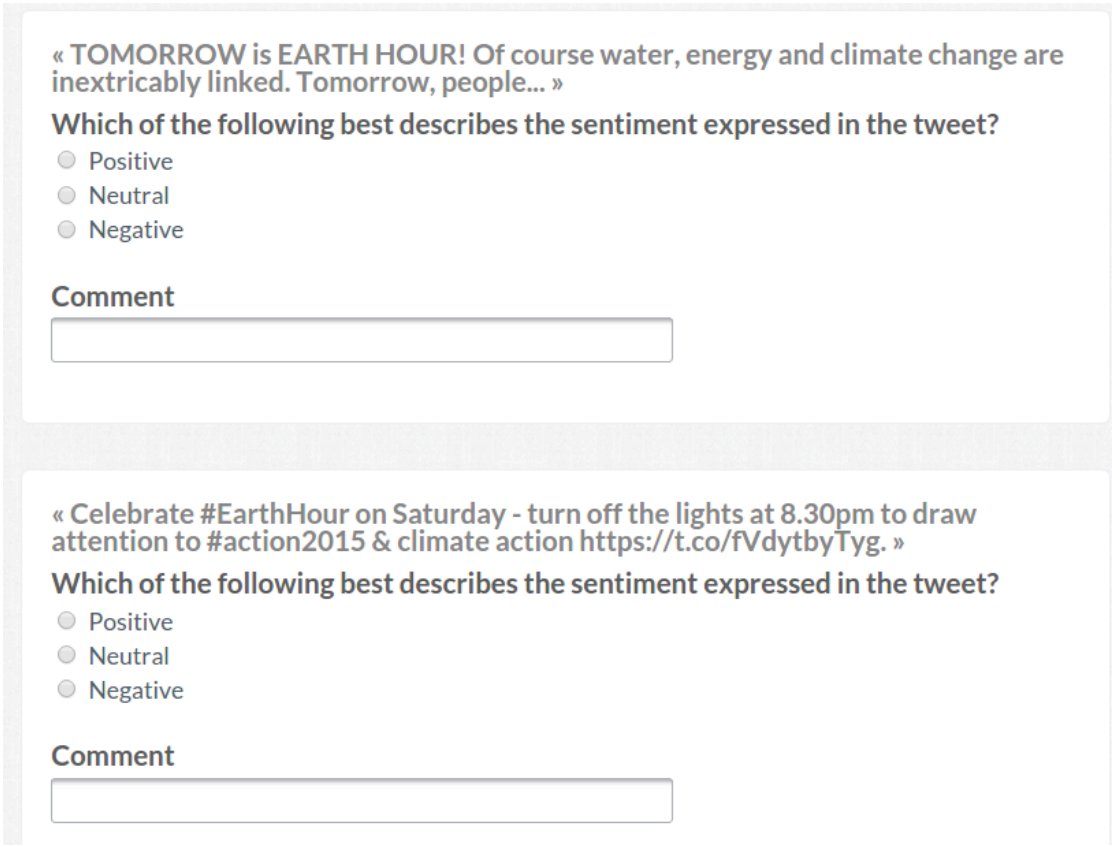
Negative:

- Earth Hour is such a stupid idea from those countries that keep empty buildings lit all night, use excessive packaging
- Totally, completly ignored the Earth Hour insanity, and I have no regrets.
- Earth Hour Day an ineffective feel good Event. Walk through your city by night...any. changes in Ligthing/Power use?

Positive:

- Show your love for the planet, and turn off your lights for #EarthHour.
- RT @tempatanfest: We are supporting PUBLIKA Earth Hour program this weekend and we're opening BDB Publika pop-up booth… https://t.co/zv1SZN….
- @TipeDarah: Happy earth hour everyone! :D http://t.co/Ei2Mv0qHKh

F Example Annotation Task in Crowdflower



DecarboNet Consortium

The Open University
Walton Hall
Milton Keynes MK7 6AA
United Kingdom
Tel: +44 1908652907
Fax: +44 1908653169
Contact person: Jane Whild
E-mail: h.alani@open.ac.uk

Waag Society
Piet Heinkade 181A
1019HC Amsterdam
The Netherlands
Tel: +31 20 557 98 14
Fax: +31 20 557 98 80
Contact person: Tom Demeyer
E-mail: tom@waag.org

MODUL University Vienna
Am Kahlenberg 1
1190 Wien
Austria
Tel: +43 1320 3555 500
Fax: +43 1320 3555 903
Contact person: Arno Scharl
E-mail: scharl@modul.ac.at

WWF Schweiz
Hohlstrasse 110
8004 Zürich
Switzerland
+41 442972344
Contact person: Christoph Meili
E-mail: Christoph.Meili@wwf.ch

University of Sheffield
Department of Computer Science
Regent Court, 211 Portobello St.
Sheffield S1 4DP
United Kingdom
Tel: +44 114 222 1930
Fax: +44 114 222 1810
Contact person: Kalina Bontcheva
E-mail:
K.Bontcheva@dcs.shef.ac.uk

Green Energy Options
Main Street, 3 St Mary's Crt
Hardwick CB23 7QS
United Kingdom
+44 1223850210
+44 1223 850 211
Contact person: Simon Anderson
E-mail:
simon@greenenergyoptions.co.uk

Wirtschaftsuniversität Wien
Welthandelsplatz 1
1020 Wien
Austria
Tel: +43 31336 4756
Fax: +43 31336 774
Contact person: Kurt Hornik
E-mail: kurt.hornik@wu.ac.at