EC Project 610829

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

D2.3.2: Environmental Opinion Extraction

28 July 2016

Version: 1.0

Version history

Version	Date	Author	Comments
0.1	20 July 2016	Diana Maynard	Initial version
1.0	28 July 2016	Diana Maynard	Final version after review

Peer reviewed by: 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 tools for environmental opinion mining delivered. Our 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 deliverable describes the second version of the tools, building on the initial version described in D2.3.1.

The report explains how to use the web service, describes the applications and the underlying natural language processing tools used, in particular focusing on the improvements made on the first version. It also details some 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, that will extend also beyond the life of the project.

Table of Contents

1. Introduction

This deliverable describes the second version of the ClimaPinion tools for environmental opinion mining in DecarboNet. It provides information about the improvements to the previous version for English opinion mining, delivered in D2.2.1, and some further evaluations. It also describes the German version of the opinion mining tool we have developed. 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. A joint paper has been published on this collaborative effort, involving work from WP1, 2 and 4 [Fernandez 2016] and an extended version is currently being prepared for a journal submission.

As with the previous version, the services require no technical skills to use, and can therefore be accessed by all project partners. 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 and json files, which means that they can be more easily integrated with other processing tools provided by the other technical partners, MODUL and OU, rather than the standard XML output normally produced by GATE and the GCP. Furthermore, the application is included in our publicly available cloud processing toolkit, known there as the DecarboNet Environmental Annotator.[1]

In this report, we first describe briefly the opinion mining web service and demo in Section 2. In Section 3, we describe the improvements to the technology underlying the opinion mining tools since the previous version. Where appropriate, we show how the performance of a particular aspect of the approach has improved since the last version. Note that some final performance evaluations on the Earth Hour datasets will be carried out in the remainder of the project, and will be described in D6.3.2. In Section 4, we describe the opinion mining tools for German we have developed and some preliminary evaluation. As for the English version, we will carry out some final performance evaluations in D6.3.2. In Section 5, we give a brief summary and outline some plans for future work which will extend beyond the life 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 (these are standard GATE components which have simply been re-used without adaptation)
- environmental term extraction as provided by the ClimaTerm tools (described in D2.2.2 and also available as a separate web service).

[1] https://cloud.gate.ac.uk/

- 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 web services are available from https://gate.ac.uk/projects/decarbonet/.

A demo is also available there, 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. The demo is improved from the previous version in two respects: first, it uses the updated methodology with better accuracy (as described in the following sections); second, it adds author and target recognition. The target recognition uses the information from ClimaTerm: if the opinion of the target is a term, a link is also provided to the ontology to which it is connected (Reegle, Gemet or DBpedia). In this example, the sentence expresses fear, which is a negative emotion. The target of the opinion is "pollution" (highlighted in yellow) and there is a link to the entry for pollution in Reegle.

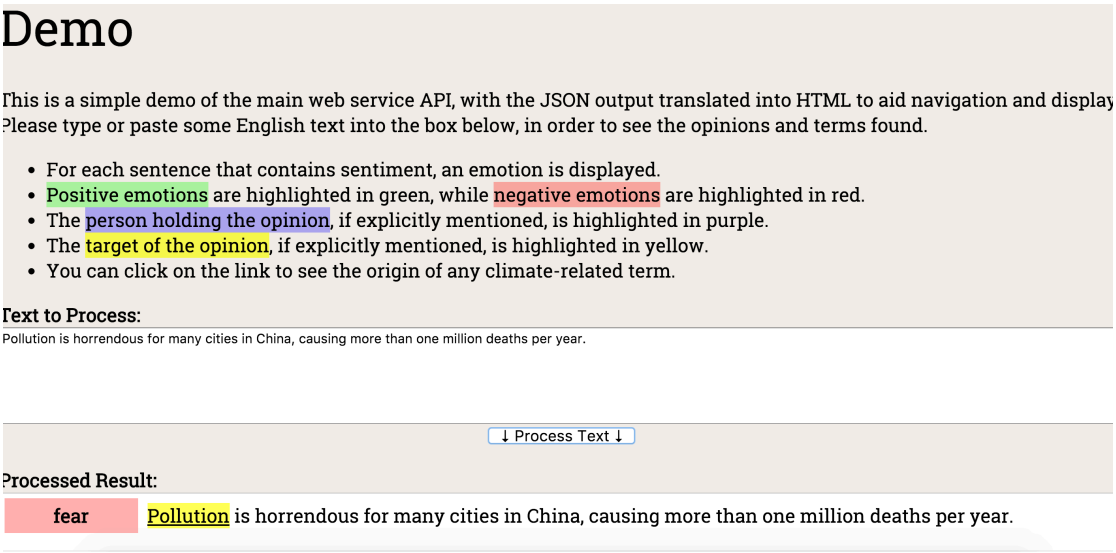

Figure 1: Screenshot of updated opinion mining demo

3. English Opinion Mining Technology

3.1. Introduction

The second version of the opinion mining tool for English, ClimaPinion, builds on the first version presented in D2.3.1, which focused mainly on high precision but to the detriment of recall. This means that the tool only detected opinionated statements where there was a fairly high degree of certainty about their accuracy. This was demonstrated in the evaluations we carried out in that deliverable, where we compared the tool with other state-of-the-art tools for opinion mining, and found that ClimaPinion produced higher Precision.

Following the error analysis conducted in D2.3.1 (depicted below in Figure 2), we have focused our attention on 3 main strands: improving the sentiment lexicons to extend their coverage but without sacrificing precision; improving some of the linguistic pre-processing components such as POS tagging; and improving the contextual relevance of the sentiment lexicons (sentiment words may change meaning in different domains).

Note that we have concentrated in this work on the error categories that were easier to find solutions for. There were some issues that cropped up frequently in the error analysis, but were difficult to resolve. For example, the "impossible" category was defined in D2.3.1 as "*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.*" Clearly this is difficult to rectify. The "language" category was defined as "*non-English tweets where our system nevertheless attempted to classify the opinion*". This is also difficult to resolve as it requires improving the language classification component. Language classification is known to be very tricky on short informal text like tweets, especially where the tweet length is very short, since these kind of classifiers need more context. The classifiers are also easily confused by the code-switching which is typical in such tweets and especially in our datasets. The following example is a tweet from our Earth Hour 2015 dataset, where the author has retweeted a tweet in English but then added some commentary in a different language, which we believe is transliterated Hindi. However, even in that non-English sentence, we find the two English words "daily" and "celebrate".

« RT @HashtagRao: "Did you celebrate earth hour?" Ji hum daily 4, 5 dafa celebrate karte hain. »

There are plans for improving the language detection component, but they will be realised in some related future projects. These plans are explained in more detail in Section 5.

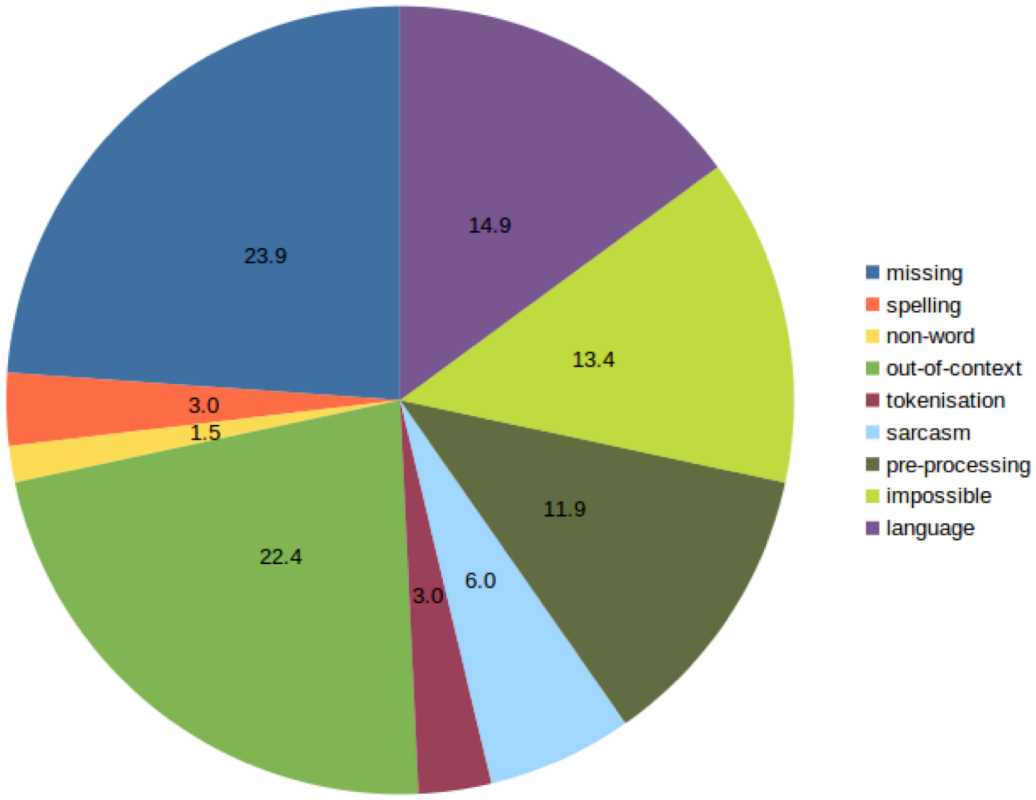

Figure 2: Error analysis of ClimaPinion v1 on Earth Hour tweets (by percentage)

3.2. Linguistic pre-processing

Incorrect POS tags impact negatively on the performance of our opinion mining tools because matches between a sentiment word or phrase from the lexicon and a word or phrase in the text are only considered valid if they both have the same part of speech. For example, the word

"like" should only be matched with a positive sentiment when used as a verb. This applies not only to sentiment-containing words but also to environmental terms which might be the target of an opinion (e.g. "lead" which is not an environmental term when used as a verb). In order to improve the POS tagging in this way, we check that all the environmental terms in our lexicons are included in the POS lexicon used by the tagger.

Another source of error is due to the Twitter-specific components such as tokenisation and normalisation. which attempt to resolve some of the issues inherent with low-quality text in social media, and to deal with phenomena such as emoticons. In particular, we discovered and fixed some errors in tokenization, such as where emoticons were incorrectly identified. Since emoticons play an important role in the detection of opinions, this was an important discovery. An example of this is the emoticon :3 (which is supposed to represent a cat's face) used to denote cuteness. However, when this is part of a date or time (e.g. 08:30) it is clearly not denoting cuteness.

We have not evaluated the impact of these pre-processing components separately, but the combination of improvements has increased accuracy by several percent on our test data (from 66.3% to 69.1% on the Earth Hour 2015 corpus described in D2.3.1 and below in Section 3.4). More evaluation of the overall performance on the Earth Hour data will be carried out during the remainder of the project and described in D6.3.2.

4. Sentiment lexicon expansion

Opinion mining on social media requires novel techniques. By its nature, social media, and particularly Twitter, exhibits new terms and orthographies, including many novel mis-spellings not found in more traditional forms of text. This increases the chance of encountering new lexicalisations of sentiment, unseen in existing gazetteers. We have experimented with three different methods for sentiment lexicon expansion, in order to improve the recall. Most existing sentiment analysis tools, especially those relying on Machine Learning techniques, use very big sentiment lexicons compiled from training on large datasets. However, the problem with these is that many of the words included in them do not directly reflect sentiment, but are simply associated frequently with sentiment-containing tweets. For instance, if "large" is more frequently associated with positive than negative tweets, it might be included in a positive sentiment lexicon, as statistically it will be a good indicator. However, this is not compatible with our philosophy of maintaining high precision, nor with our method of opinion mining based on combining sentiment-containing words with rules to assign sentiment scores based on polarity modifiers and to identify the scope of the opinion (such as opinion holder and opinion target assignment).

4.1. Thesaurus-based expansion

The first approach to automatically expanding the sentiment lexicons is based on using a thesaurus to find related words missing from the lexicon. For each emotion category, we checked to see if there were synonyms in WordNet or Roget's Thesaurus not existing already in the list. For WordNet, a plugin exists in GATE to search for a word in the thesaurus and return various features such as synonyms and hyponyms. For Roget's Thesaurus, we checked manually online[2] for synonyms of the key term in each list. Figure 1 shows some examples of synonyms found for the words cute and angry.

Table 1 shows the key term for each list, the number of terms in the original lists and the number of new terms added. In total we acquired 200 new terms with this method, although removing duplicates (as some terms were found in multiple lists) gives us 153 new terms. Experiments with the Earth Hour 2015 corpus showed, however, that this improved accuracy

[2] http://www.thesaurus.com

by only one document out of 600, and on the Earth Hour 2014 corpus it did not change the accuracy at all. One of the reasons for this is that the Earth Hour corpus is skewed heavily away from negative sentiment (only 6.9% of those tweets are negative) while the number of negative emotion words is almost double that of positive emotion words, as is the number of new emotion words added (see Table 2), so we would expect the additions to the lexicon to only have a small impact.

Table 1: Size of emotion lexicons before and after thesaurus-based expansion

List	Key Term	Original	Thesaurus	Total
Anger	Angry	259	23	282
Bad	Bad	15	38	53
Cute	Cute	12	13	25
Disgust	Disgusting	68	29	97
Fear	Afraid	162	24	186
Good	Good	14	38	52
Joy	Happy	401	19	420
Sadness	sad	214	16	230
Total		1145	200	1345

Table 2: Size of positive vs negative emotion lexicons before and after thesaurus-based expansion

Polarity	Original	Thesaurus	Total
Positive	427	70	497
Negative	718	130	848
Total	1145	200	1345

4.2. Brown Clustering

The meaning of words is often strongly related to the surrounding context. Some methods for mining word meaning exploit this distributionality by attempting to group together semantically similar words, by virtue of them having similar contexts. Brown clustering is a hierarchical agglomerative clustering technique for this, which sequentially tries to pair word types in a corpus, merging in order of minimum information loss in terms of word context. The method adopts a class-based language model, i.e. one where probabilities of words are based on the classes (clusters) of previous words. This is used to address the data sparsity problem inherent in language modelling. In simple terms, Brown clustering works by repeatedly merging the two "most similar" word classes into a single word class. Every word starts in its own class, so there are many classes with one word each. By the end, there is one class, with all the words. The order in which this happens gives the structure to the clustering. Normally, "most similar" is defined by looking at words in the left and right context, and making merges between things that have the most similar distribution.

The result is a binary tree of word types in any given dataset. Clusters can then be extracted from this tree by choosing a level of detail (i.e. a merge number) as a boundary to stop extracting at, which then leaves words in groups. This can lead to clusters that express a particular concept, which may be at a high level (e.g. time-related words) all the way down to

lexical level (e.g. variations on the word “pretty”). For example, with 1000 clusters built from a few hundred million tweets, the following cluster was generated for the word “tomorrow” [Ritter 2011]:

2m, 2ma, 2mar, 2mara, 2maro, 2marrow, 2mor, 2mora, 2moro, 2morow, 2morr, 2morro, 2morrow, 2moz, 2mr, 2mro, 2mrrw, 2mrw, 2mw, tmmrw, tmo, tmoro, tmorrow, tmoz, tmr, tmro, tmrow, tmrrow, tmrrw, tmrw, tmrww, tmw, tomaro, tomarow, tomarro, tomarrow, tomm, tommarow, tommarrow, tommoro, tommorow, tommorrow, tommorw, tommrow, tomo, tomolo, tomoro, tomorow, tomorro, tomorrw, tomoz, tomrw, tomz

Generalised Brown clustering [Derczynski 2016] is a variation which allows easy post-hoc scaling of cluster granularity through roll-up feature generation. We used this method to extract concept groups from tweets, based on an initial training dataset of 100,000 environmental tweets. While successful in some cases, it was noted that the way clusters aligned with our input sentiment gazetteers was not deterministic. For example, some clusters contained all-positive or all-negative words, while others contained a mixture of both (with clusters being about e.g. highly emotional words but not necessarily about the same emotion). Another problem was when terms were collated with other, more frequent senses of terms in the case of polysemous sentiment gazetteer entries. This suggests that a more nuanced expansion technique is required. We discuss this possibility in Section 5.

A second experiment investigated using a supervised approach. In this method, we manually merge together all the words that we think are already similar, and then let the algorithm take over after that. This means that we start with a multi-word class, that contains all of the terms we consider are related (a seed set). For this we used two gazetteer lists, one containing positive words and another containing negative words. Other terms with similar distributionality may then be attracted to this cluster, thus automatically extracting a group of

Seeds: agree best better efficient kind outstanding popular positive quality quiet soft strong success top

Cluster: lower higher strong better top positive best quality outstanding success agree

Figure 3: Example of an expanded list using Generalised Brown clustering

concordant terms. The reason for not using the emotion lists for this experiment was that with only a small corpus of 100k words, many of the emotion words would not occur frequently (if at all) in the corpus, and therefore would affect the clustering negatively [Ciosici 2015]. We therefore used the positive and negative lists, as these contained words more likely to occur in the corpus. Figure 3 shows an example of the top part of an expanded cluster of positive words, with the most frequent term occurring first. Terms in red denote errors.

With this method, we achieved some degree of success, but there was still a lot of noise, which after investigation was attributed to two major causes. First, the input dataset of 100,000 tweets (~140,000 tokens) was not large enough to show consistent distributional variations in a way that gave helpful generalisations. In fact, non-related terms that had identical context could sometimes be merged in early, diluting the meaning of the cluster and leading to some other poor choices.

Second, the point at which to trim the tree varied depending on the concept learned. Different concepts are present in the data at different levels of significance: if we imagine the language in the DecarboNet datasets as having a conceptual hierarchy (like WordNet), each gazetteer's concept occurs at a different level. This makes it hard to know where to cut.

Other issues were polysemous seeds (e.g. “quality”, which can be a positive adjective or a non-sentiment-bearing noun) which gave a mixed concept representation, and antonyms occurring in similar contexts (e.g. “higher” and “lower”). Furthermore, broad coherency in the

seed set led to high variation in the distributional context (for example, “agree” is very different from “soft”). Solutions to these problems could be using finer-grained seed sets, such as the emotion lists, along with a bigger corpus. However, the algorithm is computationally expensive, which is the main reason why we performed the initial experiments with only 100,000 tweets. Further experimentation is planned with larger sets, but first, a different method using word embeddings, which is much more efficient, was undertaken.

4.3. Word embeddings

Word embeddings are shallow, two-layer neural networks, that are trained to reconstruct linguistic contexts of words. They are based on the idea of the distributional hypothesis [Harris 1954]. The idea behind this is that semantically or syntactically similar words have similar contexts. For example, the same verb might be followed by similar kinds of nouns, e.g. “eat” is typically followed by a kind of food.

Word2Vec [Mikholov 2013] is a set of models providing word embeddings: basically an efficient implementation of the continuous bag-of-words and skip-gram architectures for computing vector representations of words. These models are shallow, two-layer neural networks, that are trained to reconstruct linguistic contexts of words, based on the idea of the distributional hypothesis. Once trained, the models can be used to map each word to a vector of typically several hundred elements, which represent that word's relationship with other words. For example, one can find the most similar words to a given term, or the most similar sentence to a given sentence. As with Brown clustering, one of the big advantages of this method is that no training data is needed, just a large corpus of relevant text. However, unlike Brown clustering, it is very computationally efficient.

We experimented with Word2Vec to see if we could extend the set of emotion words for each category (anger, fear etc.) based on their distribution in a large corpus of environmental tweets, in a similar way to the Brown clustering experiments. We performed some simple experiments to see which terms were the most similar to a given emotion word. After training on a corpus of 250k tweets about the environment, the 20 terms most similar to the word “sad” were the following:

feel_guilty, oh_shit, dumb, annoyed, bcs, home_alone, uh, ugh, wanna_go, upset, aliens, kca, fell_asleep, bomb, n’t_mean, whoops, imma, i’ve, sooo, guess.

Unfortunately, Word2Vec gives us no way of stipulating that the results should be a particular part-of-speech, so we have to post-filter these if we only want (for example) adjectives and adverbs, or if we want to separate different parts of speech, or single words from phrases. A manual process of filtering the results is necessary because there are many irrelevant terms found. However, the process is worthwhile because some good new terms are found. Note that the method also finds multi-word terms (phrases) as well as single words. For example, the manually filtered list of phrases returned as similar to words in the general list “good” was as follows:

beautiful planet, cant wait, clean energy, clean renewable, clean tech, cleaner energy. could win, create jobs, dont miss, easy way, #fabearthhourchallenge, getting excited, great excuse, great hall, great prizes, help environment, joining us, lets give, lets support, lovely planet, millions switching, much fun, noble cause, please send, positive actions, really easy, show support, showing support, simple way, something fun, spreading word, sustainable living, warm glow, well spent.

We include hashtags here even though they are technically a single smushed-together word, because they are comprised of multiple tokens. It is important to understand the benefit of this list of phrases compared with simply recognizing single positive words. There are several ways in which they can be more powerful:

- Combined words: sometimes two words do not typically denote sentiment on their own, but only when used in combination, e.g. *create jobs*, *simple way*, *warm glow*, *spreading word, cant wait.*
- Contextual relevance: when two general words are combined to form a phrase particularly relevant to the domain (and exhibiting sentiment there), e.g. *clean energy, help environment.*
- More reliable indicator: combining two words may make the sentiment indication more reliable, eg. *positive action, lets support, noble cause.*

Table 3: Size of emotion lexicons before and after embeddings-based expansion

List	Original terms	New terms	Total
Anger	259	33	292
Bad	15	39	54
Cute	12	5	17
Disgust	68	21	89
Fear	162	94	256
Good	14	22	36
Joy	401	113	514
Sadness	214	73	287
Total	1145	607	1752

Overall, we get 607 new terms from this method, as depicted in Table 3, although some are duplicates because they belong to multiple categories. Removing duplicates, we get 507 new terms. If we compare the additional terms found by this method with those found by the thesaurus method, we find for most emotion categories a greater number of new terms for the embeddings method, with the exception of *cute, good,* and *disgust.* For a fair comparison, however, we should remove the phrases from the embeddings lexicon, since the thesaurus method did not generate phrases. Table 4 shows a comparison for the single word terms only, and the overlap between words generated by the two methods (i.e. how many were generated by both the thesaurus method and the embeddings method). We can see clearly that the overlap is very small, which means that combining the two methods is potentially worthwhile.

Let us give some examples to illustrate the process. For the original list “cute”, which contained 12 words, 7 of these were found in the training corpus, and out of the 140 words found for these via the embeddings method, only 6 were deemed relevant. Note that some of those 140 words were duplicates, however, because the top 20 most similar terms were calculated separately for each of the 7 original words.

An example of a larger list is the one for “anger”. This contained 265 words, of which 19 were found in the training corpus, and from which 33 new terms were generated. Typically, adjectives were found in the corpus more often than nouns and verbs (e.g. “angry” but not “anger”, and unsurprisingly, more formal words such as “irate” were less likely to be found.

Table 4: Comparison between thesaurus and embeddings method (single-word terms only)

List	Thesaurus	Embeddings	Overlap
Anger	23	33	7

Bad	38	35	0
Cute	13	5	1
Disgust	29	20	4
Fear	24	79	0
Good	38	22	4
Joy	19	55	1
Sadness	16	58	3
Total	200	607	20

In experiments on the Earth Hour 2015 corpus, by adding these new terms to our sentiment lexicons, we improved both Precision and Recall, increasing the accuracy by almost 1%. Admittedly, the improvement is small (we essentially corrected 3 sentiment polarity detection errors in the corpus). The additional lexicons (from the embeddings experiment) led to finding an extra 85 mentions of emotion words in the corpus. This suggests that while they may not have brought much additional information in this scenario, possibly because enough emotion information was already found in the tweet (e.g. another positive emotion word was already found), they could be useful in other cases. Further experimentation is needed to investigate in more detail why the additional words were not as helpful as we might have expected, however.

We plan to experiment with training also on non-environmental tweets, which would give us a much bigger training corpus that might also be both more balanced and richer in terms of emotion words. On the other hand, this might be detrimental as it could contain words which are used differently in the environmental domain.

5. Evaluation

We have re-evaluated the tools as they currently stand after the combined improvements, on the Earth Hour 2015 dataset, since this is probably the most reliable (being annotated by crowdsourced workers and triple-annotated with adjudication) and relevant (since the tweets refer directly to one of our case studies in the project, carried out in WP6. Table 5 shows the evaluation results reported in D2.3.1 for the first version of the ClimaPinion tool, compared with 3 other state-of-the-art tools. From this evaluation, we found that Recall of sentiment-containing tweets was quite low: the system often annotated a tweet as having neutral sentiment instead of positive, because it did not find relevant sentiment. In Table 6, we show the overall improvement from version 1 to version 2.

Table 5: Performance of different opinion mining tools on the Earth Hour 2015 dataset

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: Performance of ClimaPinion v1 and v2 on the Earth Hour 2015 dataset

Tool	Correct	Incorrect	Accuracy
ClimaPinion v1	398	202	66.33 %
ClimaPinion v2	436	164	72.67 %

6. German opinion mining

We have developed the first version of a ClimaPinion tool for German by taking a very basic existing German opinion mining tool developed in GATE during the ARCOMEM project [Maynard 2015] and incorporating a number of improvements. The German ClimaPinion tool is similar to the English tool, but almost all of its component processing resources are adapted to German. Specifically, it uses the following components:

- a set of German pre-processing resources (tokenization, sentence splitting etc.);
- a German POS tagger and Named Entity tagger (newly incorporated in GATE);
- German sentiment lexicons originating from the German version of SentiWordNet (improved version from ARCOMEM);
- Components for finding specific linguistic constructs in German (adapted from English ClimaPinion components);
- German term recognition component ClimaTerm (described in D2.2.2);
- German grammar rules for finding sentiment and emotions and annotating sentiment, authors and targets (improved version from ARCOMEM).

Figure 4 shows a simple example of a tweet annotated by the German ClimaPinion in GATE. The tweet originates from a collection of German tweets downloaded from the Media Watch for Climate Change and used for experimentation and testing. The tweet can be translated into English as "*The manipulated VW-Diesels are so toxic*." The context of this tweet was the recent scandal where it was discovered that Volkswagen had intentionally programmed some of their diesel engines to activate certain emissions controls only during laboratory emissions testing. Here the tweet correctly shows negative polarity, and an angry emotion. In the rest of this section, we describe the various components which make up the application, and explain how they are combined.

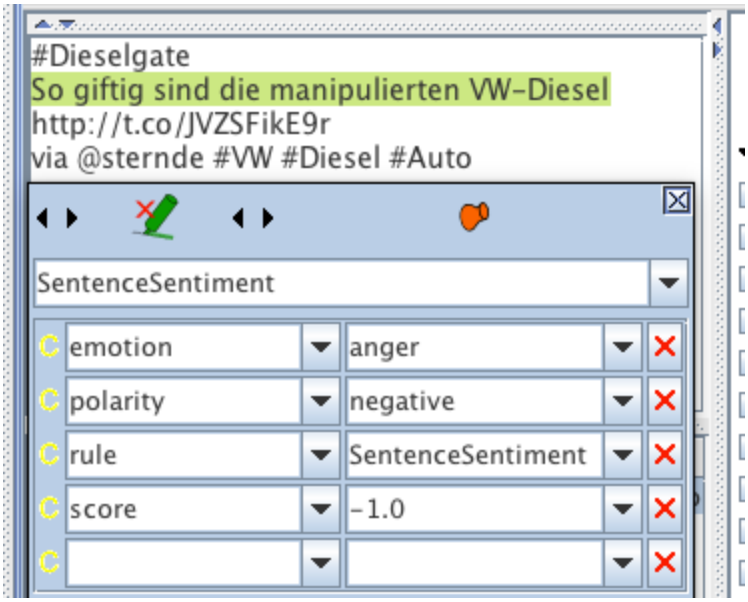

Figure 4: Screenshot of German opinion mining in GATE

7. Components of the German opinion mining tool

We have built on the basic German opinion mining tools in GATE in two main ways. First, we have made some specific adaptations to deal with the task and domain: this concerns the detection of opinion targets (the term or entity that the opinion is about), 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.

8. Linguistic pre-processing and term recognition

The linguistic pre-processing components are essentially the same ones as used for the German ClimaTerm, described in D2.2.2. We use the universal tokeniser and sentence splitter included in ANNIE, GATE's standard English Information Extraction plugin, and which are also used in GATE's standard German Information Extraction plugin, available as part of GATE. We swap the English POS tagger and Named Entity tagger in ANNIE for a German-specific tagger based on training models from Stanford CoreNLP[3]. We also directly reuse the German ClimaTerm tool (described in D2.2.2) to find environmental terms (this is required for the opinion target detection, as for English).

9. Sentiment lexicons

The sentiment lexicons are originally derived from the German version of SentiWordNet [Remus 2010], but have been augmented by us with additional terms. Some of these were added manually after experimenting on training data. In total these comprise 3468 terms, split almost equally between positive and negative ones. A more comprehensive set of terms was added by translating the lists of emotion terms from the English version into German, which adds another 1340 terms to the gazetteer (though there is some overlap). This means that we can also categorise the sentiment into different emotions (fear, anger, joy, etc. as for the English ClimaPinion). We included in the translated list the synonyms generated in the English version and described in the previous section. Due to the overlap between emotionally categorized terms and those categorized just as positive/negative, in our grammar rules we always prefer to make use of an emotion term over a non-emotion sentiment term. Any word that does not have an explicit emotion category gets simply labelled as positive or negative as the value of the emotion.

10. Recognition grammars

The recognition grammars are also adapted slightly from the English ones, due to the use of a different tagset for the German parts-of-speech (TIGER[4] as opposed to the Penn TreeBank) and due to some differences in the way German terms may be formed. They are substantially improved from those in the original German opinion mining tool from ARCOMEM, which did little more than recognize the sentiment words from gazetteer lists.

[3] http://stanfordnlp.github.io/CoreNLP/

[4] http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html

The new grammars first check that the part-of-speech category on the word in the text matches the part-of-speech category on the word in the lexicon before a match is made. This is because words can have different meanings (and different polarity) when used as different parts of speech. For example, in English, the word *"kind"* as an adjective would be viewed as positive, but as a noun it does not indicate sentiment. So we only want to match the lexical entry for it if it is used as an adjective in the text.

The grammars also recognise questions, since we typically do not want to attach sentiment to a question. For example, the question "*Are you happy?"* does not convey sentiment. We then have some options for modifying the score and polarity according to other linguistic elements. For example, a sentiment-containing adjective may have its score altered according to the context: modifiers such as adverbs may strengthen the score, negation may reverse the score from positive to negative or vice versa, swear words may also strengthen the score, and so on. A number of rules deal with these issues, as for the English version of ClimaPinion. Sentiment is output at various levels – for each sentence, an annotation is created if positive or negative sentiment is present, detailing the polarity (positive or negative), score (strength of sentiment from -1 to +1), and emotion. For each tweet, the sentiments are then aggregated as for the English version, by aggregating the scores and calculating the mean. Emotion is not aggregated because there is no obvious way to do this. We can thus look at sentiment either at the sentence level or at the tweet level.

As for the English version, sentiment targets are also annotated, i.e. what the sentiment is about. The sentiment target is restricted currently to environmental terms and entities. For example, in the following German tweet: "*Klimawandel ist mitverantwortlich für den starken Rückgang an Hummelarten in Europa und Nordamerika.*" (*Climate change is partly responsible for the sharp decline in bumblebee species in Europe and North America.*), the sentiment is annotated as negative and the target is climate change, because the author of the tweet is expressing a negative opinion about climate change (that it is destroying bumblebees). Tweets which have sentiment but no explicit target are also annotated as a general positive or negative sentence. For example, the tweet "*Nicht mehr ganz neu*" (*Not really new any more*) conveys generally negative sentiment but does mention what the opinion is about.

11. Evaluation

Some preliminary evaluation has been carried out by testing our ClimaPinion tool on a very small sample German dataset of annotated opinions kindly provided to us by the uComp project.[5] This dataset is very richly annotated with 17 fine-grained kinds of emotion and opinion, as well as many other related classes. They divide opinions into three types, each of which has several subtypes, and also distinguish between emotion, sentiment and opinion, as depicted in Table 7. A more complex description of their categorization is given in [Fraisse 2014]. For comparing opinion polarity, we can simply use the coarse-grained positive and negative categorization, depicted in the table by green and red colours respectively, as the basis for our evaluation. For comparing emotion categories, we suggest the mapping shown in Table 8. We do not map the uComp opinion categories as they seem a little out of scope in our context.

Emotion		Sentiment		Opinion	
displeasure	pleasure	dissatisfaction	satisfaction	devaluing	valuing

[5] http://www.ucomp.eu/

disturbance	appeasement			disagreement	agreement
contempt	love				
surprise	surprise				
anger	fear				
sadness	boredom				

Table 7: uComp opinion categories

ClimaPinion	uComp
Anger	Anger Disturbance
Bad	Displeasure Boredom Dissatisfaction Negative Surprise
Cute	
Disgust	Contempt
Fear	Fear
Good	Appeasement Pleasure Satisfaction Positive Surprise
Joy	Love
Sadness	Sadness

Table 8: Mapping between ClimaPinion and uComp emotion categories

Currently, this uComp dataset provided to us is only a very small sample, containing 23 tweets, although evaluation on a larger set is planned once this has been finalized. The larger dataset will be used as training material for an upcoming public evaluation organized jointly between us and other uComp project partners. On this dataset, using the mappings described above, our tool obtained 86.9% accuracy on classification of tweets into positive, negative and neutral. However, this result should be taken with a little caution, not only because of the small size, but also because the gold standard annotation on this dataset was performed in such a way that the sentiment judgements were made on small sections of the tweet and not over the tweet as a whole. We then aggregated these over the tweet itself, but this aggregation process may not accurately reflect the sentiment of the tweet itself. We can view this in a similar way to a baseline sentiment analysis tool which just takes into account the lexical polarity but not sentence structure which could change the polarity of the tweet (sarcasm, conditional sentences, questions etc). We can therefore only use this evaluation to give us a preliminary notion of success and to indicate that we are on the right track.

We also compared the emotion values (after mapping) between our tool and the uComp annotations, but we found too many discrepancies to read much into it. For example, the guidelines for annotation were rather different (span boundaries of annotated elements were

different, plus many words annotated by us were not annotated in the uComp dataset, for unknown reasons, so this experiment was not particularly useful. Again, this stems largely from the fact that the emotion annotations in the uComp set were on individual words rather than on sentences or tweets as a whole. We will therefore investigate evaluation specifically on our own environmental collection during the remainder of the project, in addition to this dataset.

12. Summary and further work

In this deliverable we have described the second 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 via GATE Cloud, and via our GCP tool which enables large-scale processing in a format easily accessible to users (csv input and output). As mentioned previously, further evaluations will be carried out in the remainder of the project duration on the Earth Hour datasets as part of WP6, and will be reported in D6.3.2.

There are a number of avenues for further research, some of which are already being carried out in related projects such as COMRADES[6] and SoBigData[7], and others which may be carried out in forthcoming ones.

In Section 3, we addressed previous shortcomings in opinion mining accuracy by focusing primarily on expanding the opinion lexicons to increase recall without sacrificing precision. The first method we experimented with, using synonyms and related words from thesauri, was successful but only mildly influential: it did not return many new words, although those that it returned were generally good.

The second method we investigated, Brown clustering, was more experimental and less successful, though still potentially promising as a strategy. Further work would be needed to see if good results can be achieved, such as experimenting with bigger datasets but also modifying the technique used to find a clustering metric that permits group expansion. Our experiments showed that distributional clusters do not always expand in a way that is conducive to building lexicons. Brown clustering is guided by mutual information, which models distributional similarity, i.e. how similar two words are based on their neighbours. Words that have similar neighbours thus have similar distributions. The clustering technique progressively merges together the two words -- or word groups -- in a way that causes the least possible loss of aggregate mutual information (mutual information across the whole input corpus). However, due to the metric used, highly frequent words tend to attract very low frequency words that have marginally similar distributions. For example, *good* might merge with *`dainty*, *profitable* and other infrequent terms, but might not merge with *excellent* because doing so would cause too much of a mutual information drop across the corpus. Therefore, to improve the performance of this kind of clustering, the clustering goal – minimized aggregate mutual information loss -- needs to be changed, so that merges of frequent and similar words are more acceptable.

The third method, word embeddings, was the most promising of the three, though also still a little limited. While it produced many new terms, the downside to this was that a significant proportion of them were irrelevant or incorrect (as additions to an existing emotion lexicon), so a substantial amount of manual effort was necessary to validate the lists and extract only the relevant ones, and this is also subjective. However, as with Brown clustering, it has the advantage of requiring no training data other than a large unannotated corpus of relevant text, and is a lot more efficient (processing takes seconds rather than hours). This means that it is

[6] http://www.comrades-project.eu

[7] http://www.sobigdata.eu

useful as a domain adaptation tool since new terms generated are likely to be domain relevant. This is particularly important when words have different meanings (and sentiment) in different domains.

As discussed in Section 3, some of the problems with the opinion mining accuracy stem from pre-processing errors such as language identification. This is also a potential problem for term recognition components and for the work in WP6 which builds on the ClimaTerm and ClimaPinion tools. For the language identification task, ideas for improvement are being examined in the EU COMRADES project. The core idea for this task is on the basis that classifying a text as a single language is different from discriminating between many languages. There are formulations that do this in general settings, e.g. from the world of outlier detection, where the idea is to identify members of the "outgroup"; this can be achieved with methods such as one-class SVM. Setting a confidence or decision boundary works directly to balance the precision and recall of the system. Extraction of good predictive features for language identification across multiple domains can be performed by examining the difference in information gain of each feature with language and with the source domain [Lui 2011], which works excellently across languages and genres, or by using a non-discrete distributional measure like char2vec (see e.g. [Kim 2016]). Another potential avenue for research could be in the metadata that typically accompanies social media content, such as the area of origin, the user's name and profile text, the top-level domains of URLs they mention; as well as the time of day. For example, we might be more likely to choose a Cebuano classification in an edge case if the document were written during peak activity times for the Philippines (e.g. afternoon / early evening in that time zone).

In DecarboNet we have focused on adapting our existing basic opinion mining tools to the environmental domain, as well as extending the functionality and improving the performance. This adaptation work is being continued in the SoBigData project, where the opinion mining tools are currently being used for analysis of Brexit tweets in the political domain[8], and were also used for analysis of the UK elections in the Nesta Political Futures Tracker[9] [Dietzel 2014, Maynard2015b]. Plans are underway for extensions also to this work in new projects, which will involve further adaptation and improvements to the opinion mining tools, such as adaptation to more languages, and improvements to the German tool to bring it further into line with the English tool in terms of performance.

[8] http://gate4ugc.blogspot.co.uk/2016/06/introducing-brexit-analyser-real-time.html

[9] https://gate.ac.uk/projects/pft/

Table 2. Typology of frames applicable to climate change

Frame	Defines science-related issue as . . .
Social progress	A means of improving quality of life or solving problems; alternative interpretation as a way to be in harmony with nature instead of mastering it.
Economic development and competitiveness	An economic investment; market benefit or risk; or a point of local, national, or global competitiveness.
Morality and ethics	A matter of right or wrong; or of respect or disrespect for limits, thresholds, or boundaries.
Scientific and technical uncertainty	A matter of expert understanding or consensus; a debate over what is known versus unknown; or peer-reviewed, confirmed knowledge versus hype or alarmism.
Pandora's box/Frankenstein's monster/runaway science	A need for precaution or action in face of possible catastrophe and out-of-control consequences; or alternatively as fatalism, where there is no way to avoid the consequences or chosen path.
Public accountability and governance	Research or policy either in the public interest or serving special interests, emphasizing issues of control, transparency, participation, responsiveness, or ownership; or debate over proper use of science and expertise in decisionmaking ("politicization").
Middle way/alternative path	A third way between conflicting or polarized views or options.
Conflict and strategy	A game among elites, such as who is winning or losing the debate; or a battle of personalities or groups (usually a journalist-driven interpretation).

SOURCES: W. A. Gamson and A. Modigliani, "Media Discourse and Public Opinion on Nuclear Power: A Constructionist Approach," *American Journal of Sociology* 95, no. 1 (1989): 1–37; U. Dahinden, "Biotechnology in Switzerland: Frames in a Heated Debate," *Science Communication* 24, no. 2 (2002): 184–97; J. Durant, M. W. Bauer, and G. Gaskell, *Biotechnology in the Public Sphere: A European Sourcebook* (Lansing, MI: Michigan State University Press, 1998); M. C. Nisbet and B. V. Lewenstein, "Biotechnology and the American Media: The Policy Process and the Elite Press, 1970 to 1999," *Science Communication* 23, no. 4 (2002): 359–91; and M. C. Nisbet, "Framing Science: A New Paradigm in Public Engagement," in L. Kahlor and P. Stout, eds., *Understanding Science: New Agendas in Science Communication* (New York: Taylor & Francis, in press, 2009).

Figure 5: Nisbet's typology of climate change frames (taken from [Nisbet 2009])

In the wider context of the application of the opinion mining tools to understand behaviour, which is discusssed further in WP6, one issue that has not been addressed is the different ways in which organisations in particular might try to influence people's opinions, often known as "framing". This idea is discussed in detail by [Nisbet 2009]:

"Reframing the relevance of climate change in ways that connect to a broader coalition of Americans -- and repeatedly communicating these new meanings through a variety of trusted media sources and opinion leaders - can generate the level of public engagement required for policy action. Successfully reframing climate change means remaining true to the underlying science of the issue, while applying research from communication and other fields to tailor messages to the existing attitudes, values, and perceptions of different audiences, making the

complex policy debate understandable, relevant, and personally important. This approach to public outreach, however, will require a more careful understanding of U.S. citizens' views of climate change as well as a reexamination of the assumptions that have traditionally informed climate change communication efforts."

Table 9: Painter's set of 4 climate change frames [Painter 2014]

Frame	Description
Disaster	Mention of possible adverse impacts or effects such as sea-level rises, more floods, water or food shortages, population displacements, damage to the coral reefs, diminishing ice sheets, etc.
Uncertainty	Mention of uncertainties about climate science, such as ranges in projections for temperature increases, sea-level rises, the possible adverse impacts, and so on. It can also be indicated by mention of the shortcomings of computer models or the presence of sceptic voices.
Opportunity	1. Opportunities accruing from doing something to reduce the risks from greenhouse gas emissions (the advantages of any move to a low-carbon economy) 2. Opportunities accruing from doing nothing and allowing climate change to take place (such as longer growing seasons in the northern hemisphere, or the prospects of new shipping routes and the possibility of mineral, gas, and oil exploration in the Arctic).
Risk	Where the word 'risk' is used, or where the odds, probabilities, or chance of something adverse happening were given, or where everyday concepts or language relating to insurance, betting, or the precautionary principle were included. A 'risk management' approach to the climate challenge would also be a strong indicator of this frame.

It is possible that we could try to map emotions displayed in opinionated tweets to different frames in the climate change context. Nisbet provides a typology of frames which could be applicable to climate change, shown in Figure 5. However, identifying this kind of frame goes far beyond the scope of the language analysis work in this project, since it demands advanced topic extraction and linguistic analysis, and would entail months of work.

Another interesting piece of work carried out around the idea of framing was an investigation of the television reports from different channels and in different countries about three IPCC Working Group reports released in 2013 and 2014 [Painter 2014]. They applied 4 frames (depicted in Table 9), representing disaster (in the sense of adverse impacts), uncertainty, explicit risk, and opportunity, to the analysis of the television bulletins, each frame designed to represent a way of representing climate change. For example, doom-laden depictions are very frequent, but are typically not conducive to personal engagement, as we also found in our early DecarboNet experiments [Fernandez 2015]. According to Painter et al., on the other hand, uncertainty can be a hindrance to decision-making, and is sometimes misunderstood as ignorance. Risk and opportunity are both often viewed as helpful in terms of encouraging engagement, although this depends on the exact situation. While these kinds of frames are very useful to study, they again go beyond the scope of the planned work in DecarboNet, since it is not evident how to map emotions or the results of other linguistic analysis to such things, and would require extensive work. Indeed, this exact work is planned in a new Grantham Centre-funded project[10] starting in October 2016, which will investigate framing in

[10] http://grantham.sheffield.ac.uk/

the context of the depiction of natural disasters by the news media and their relation to climate change. This work will build on all the language analysis tools developed in DecarboNet.

13. A. List of Figures

14. B. List of Tables

15. 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

16. D. References

M. Ciosici, 2015. Improving Quality of Hierarchical Clustering for Large Data Series. Masters' Thesis, Aarhus University, Denmark.

L. Derczynski and S. Chester, 2016. Generalised Brown Clustering and Roll-up Feature Generation. In Proc. 30th AAAI, Phoenix, AZ USA.

M. Fernandez, G. Burel, H. Alani, L. Piccolo, C. Meili, and R. Hess (2015). Analysing engagement towards the 2014 Earth Hour Campaign in Twitter. In: EnviroInfo &ICT4S 2015: Building the Knowledge Base for Environmental Action and Sustainability, 7-9 September 2015, Copenhagen, Denmark.

M. Fernandez, H. Alani, L. Piccolo, C. Meili, D. Maynard and M. Wippoo, 2016. Talking Climate Change via Social Media: Communication, Engagement and Behaviour. Proc. of WebSci, May 22-25 2016, Hannover, Germany.

A. Fraisse and P. Paroubek, 2014. Toward a Unifying Model for Opinion, Sentiment and Emotion Annotation and Information Extraction. 9th Language Resources and Evaluation Conference (LREC-2014). Reykjavik, Iceland.

Y. Kim, Y. Jernite, D. Sontag and A.M. Rush, 2015. Character-aware neural language models. In Proc. AAAI 2016. arXiv preprint arXiv:1508.06615.

Lui, M. and Baldwin, T., 2011. Cross-domain feature selection for language identification. In In Proceedings of 5th International Joint Conference on Natural Language Processing.

D. Maynard and J. Hare, 2015. 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.

Nisbet, M.C., 2009. Communicating climate change: Why frames matter for public engagement. Environment: Science and Policy for Sustainable Development, 51(2), pp.12-23.

Painter, J., 2014. Disaster averted? Television coverage of the 2013/14 IPCC's climate change reports.

Remus, R., Quasthoff, U. and Heyer, G., 2010. SentiWS - A Publicly Available German-language Resource for Sentiment Analysis. In Proc. of LREC 2010.

A. Ritter, S. Clark, and O. Etzioni, 2011. "Named entity recognition in tweets: an experimental study." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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