Challenges of Evaluating Sentiment Analysis Tools on Social Media

Diana Maynard, Kalina Bontcheva

University of Sheffield, Department of Computer Science
Regent Court, 211 Portobello, Sheffield, S1 4DP, UK
{d.maynard,k.bontcheva}@sheffield.ac.uk

Abstract

This paper discusses the challenges in carrying out fair comparative evaluations of sentiment analysis systems. Firstly, these are due to differences in corpus annotation guidelines and sentiment class distribution. Secondly, different systems often make different assumptions about how to interpret certain statements, e.g. tweets with URLs. In order to study the impact of these on evaluation results, this paper focuses on tweet sentiment analysis in particular. One existing and two newly created corpora are used, and the performance of four different sentiment analysis systems is reported; we make our annotated datasets and sentiment analysis applications publicly available. We see considerable variations in results across the different corpora, which calls into question the validity of many existing annotated datasets and evaluations, and we make some observations about both the systems and the datasets as a result.

Keywords: evaluation, sentiment analysis, social media, annotation

1. Introduction

In the last decade, there has been a plethora of research on various forms of sentiment analysis (Pang and Lee, 2008), from tools to analyse product reviews through to more complex tasks such as understanding and predicting political voting from social media. Indeed, the terms opinion mining and sentiment analysis are now used interchangeably, and can be used to cover opinion detection and classification, emotion classification, opinion reliability, opinion stance, strength of opinion, detection of opinion holders and targets, and more. However, fair evaluation of such systems is fraught with difficulty, partly because the task is often subjective (human annotators do not agree on what is correct), and partly because comparing systems fairly is complicated when they tackle the problem in different ways. In this paper, we discuss some of the typical problems with comparing sentiment analysis tools, creating annotated corpora and interpreting the results in a meaningful way, and study the impact of this on 3 different corpora.

The DecarboNet project[1] aims at understanding the potential of social platforms in mitigating climate change. Within the project, we have developed tools for analysing environmental tweets with respect to the opinions expressed and topics discussed, investigating correlations between social media engagement and behavioural change (Maynard and Bontcheva, 2015; Dietzel and Maynard, 2015). We have evaluated the tools in a case study based around the Earth Hour events (Fernández et al., 2015), by comparing with state of the art tools on both an existing twitter corpus and a domain-specific crowdsourced evaluation corpus we have created. The datasets are available from our DecarboNet project pages[2]. These efforts have motivated the discussion in the rest of this paper.

2. Creation of a manually annotated sentiment corpus: Earth Hour 2015

Earth Hour is an annual global event where people switch off their lights for one hour to show they care about the future of the planet. We created a twitter corpus by downloading all tweets in English about Earth Hour 2015, and selecting at random 600 of them. This corpus was then annotated manually for sentiment, and is publicly available. Using GATE's crowdsourcing plugin (Bontcheva et al., 2014), we assigned the dataset to 16 annotators, such that each tweet was triple-annotated. The crowdsourcing plugin offers infrastructural support for mapping documents to crowdsourcing units in CrowdFlower and back, as well as automatically generating reusable crowdsourcing interfaces for NLP classification and selection tasks. Essentially, it provides a workflow enabling users to pre-annotate documents with linguistic units, export the documents to CrowdFlower and set up the task, and then import the resulting annotated documents back into GATE if needed, where manual or automatic adjudication can then be performed.

Each person annotated between 50-200 tweets: the maximum was set at 200 to prevent the set becoming too biased by a single annotator and so that annotators would not become bored and make mistakes. The latter is a common problem when large annotated corpora are developed. The annotators were all fluent in English and had a good understanding both of the task and of climate change and Earth Hour. The instructions given to them are shown in Figure 1, and were designed to be succinct but clear: we tried to minimise ambiguity by instructing the annotators to use neutral if they were not sure about polarity. This was in line with the decisions made by the system, but is actually not typical of sentiment analysis tools or corpora, and could account for bias towards our system when comparing with other tools.

The GATE crowdsourcing plugin 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 these were easily resolvable.

Although in general, the annotators found the task quite easy (according to verbal feedback), it was sometimes not

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

[2] https://gate.ac.uk/projects/decarbonet/

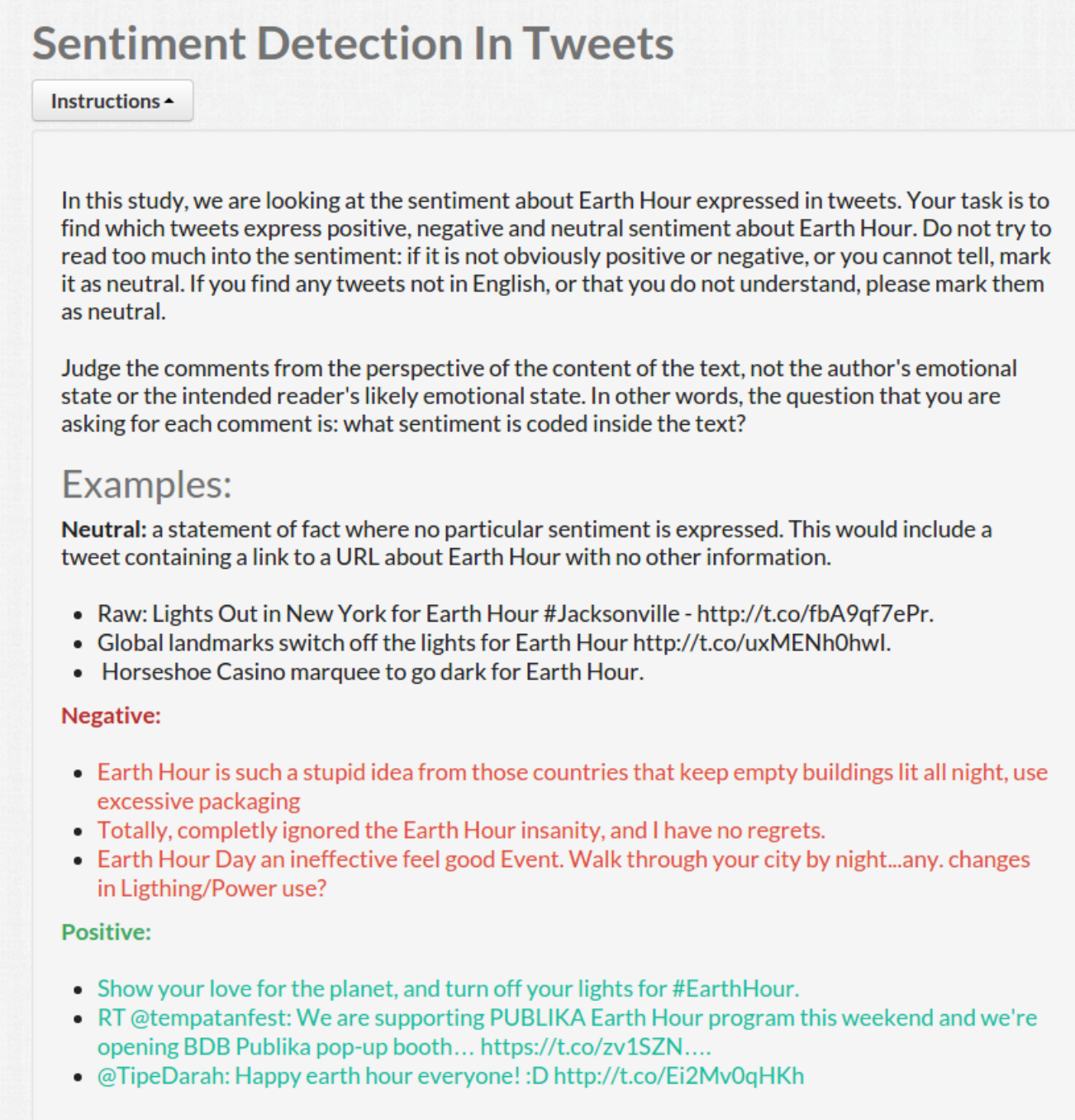

Figure 1: Instructions given to annotators

clear to them 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 deemed it neutral.

Inter-annotator agreement was measured using Fleiss' kappa, with a score of 44.19. There is no generally agreed measure of significance for this; according to (Landis and Koch, 1977) our score indicates moderate agreement, though this is by no means universally accepted, and the number of categories does affect this score. While the kappa score is quite low, we do use the majority judgement on the tweets, so where one out of three annotators disagreed is not so important. It does, however, emphasise the difficulty of the task.

The proportion of judgements is interesting: positive and neutral were much more frequent than negative. 47.83% of tweets were neutral, while 45.28% were positive, and only 6.89% negative. We have found this to be typical with tweets about Earth Hour, because people posting about it are either simply informing, or are sharing positively; those who do not care about Earth Hour typically are not concerned with tweeting about it. There are, however, some negative tweets about Earth Hour, posted by those who

Positive	– Show your love for the planet, and turn off your lights for #EarthHour – @getflipp Earth hour and Earth day will be so welcomed! love the power of GREEN! – Tomorrow, unplug for #EarthHour - these polar bears will thank you! http://t.co/NITs4YynAM.
Negative	– Earth Hour based on a myth http://t.co/qCChRLghHP via @joTurkishWeekly. – RT @hockeyschtick1: Earth Hour - the hour when warmists prove they have absolutely no self-awareness. – RT @PaulHsieh: Not doing "Earth Hour": I think our ER patients would prefer our CT and MRI scanners were powered up on a Saturday night!
Neutral	– Why doesn't earth hour correspond with earth day? – Earth Hour: Everything you need to know about the global event - ChronicleLive http://t.co/ZB42GfcIHg. – 'Earth Hour' to be observed tomorrow. http://t.co/d8ifRaecgf

Table 1: Examples of tweets from the Earth Hour 2015 corpus

think that such an activity is a waste of time, or that bad things might happen when people turn their lights off (increased crime, for example). Table 1 shows some examples of positive, negative and neutral tweets from the corpus.

The predominance of positive and neutral tweets in the corpus begs the question of whether an evaluation set should reflect real life or should be more balanced in order to test a system more thoroughly. For example, it turned out, as discussed later in Section 4., that our tools were worst at handling negative tweets, but since the proportion of negative tweets was extremely low in our corpus, it did not affect the results too badly. One can argue convincingly that this is a valid approach to take if the real life datasets for which the tool will be used also exhibit the same skewed nature, but it may account for differences in performance levels on other corpora.

3. Problems with existing annotated corpora

Using existing annotated corpora for an evaluation is not always straightforward, because different real life tasks involving the same corpus may require different solutions, and the designers of the task often have different ideas about what constitutes a correct solution. (Reckman et al., 2013) describe the difficulties understanding the development dataset used for the Sem-Eval 2013 Task 2, where target terms must be annotated with positive or negative polarity independently from the sentiment of the sentence as a whole. They found it (unsurprisingly) unusual to label words such as *like* as positive when they occurred as part of longer negative phrases such as *I didn't like*. Depending on how a system treats negative expressions, this may be tricky to break down into smaller segments, e.g. if a system uses phrases pre-annotated with sentiment, rather than individual words. Second, they found that it was not clear when words such as *apologise* should be treated as positive and negative, as these were annotated inconsistently in the gold standard. This also brings a wider point – even when such words are consistently annotated within a single corpus or evaluation, they may be annotated completely differently in another one with different guidelines. This makes it very hard to compare systems trained on different datasets or developed using different ground truths.

Another issue involves the distinction between neutral and no sentiment. Some systems distinguish between these; neutral being used where there is an equal number of positive and negative elements or where the author clearly is expressing some sentiment but it is unclear exactly what. However, since both manual annotators and automated tools often struggle to distinguish between these two cases, we (and others) use neutral and no sentiment interchangeably.

A final major shortcoming of many existing evaluation datasets is the lack of specifications provided about the annotation methodology (Saif et al., 2013). For example, (Go et al., 2009) do not report the number of annotators in the commonly used Stanford Twitter sentiment (STS) corpus[3], while many datasets do not provide information about inter-annotator agreement. Single annotated corpora can be treated quite suspiciously, since they are so prone to bias, especially when annotated by the developer of the tool evaluated on it, as is often the case. Indeed, this is one of the reasons why we compare a single-annotated corpus (Earth Hour 2014) with a crowdsourced triple-annotated and adjudicated corpus (Earth Hour 2015) in this work, as described in the following section.

4. Experiments

We performed a set of experiments to compare 4 different sentiment analysis tools on 3 corpora: the SentiStrength twitter corpus[4], the Earth Hour 2015 corpus described above, and a smaller corpus similar to Earth Hour 2015 but annotated by a single user (the developer) and comprising tweets about Earth Hour 2014. We compared 4 sentiment analysis systems on these: SentiStrength (with default settings); the rule-based ClimaPinion system developed in GATE specifically for analysing environmental tweets; an older GATE-based general domain system (ARCOMEM); and a lexicon-based system developed for the DIVINE project (Gindl et al., 2010). The choice of these systems was motivated by their easy availability and similarity of method (being strongly lexicon-based), which nevertheless offers some rather diverse results. Results for accuracy are shown in Table 2, where SS stands for SentiStrength (corpus) and EH stands for Earth Hour (corpus). The three systems against which we compare ClimaPinion 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

[3] http://help.sentiment140.com/

[4] available at http://sentistrength.wlv.ac.uk

Tool	SS	EH2014	EH2015
SentiStrength	59.17	66.20	65.00
ClimaPinion	57.21	86.80	66.33
ARCOMEM	46.04	70.34	47.83
DIVINE	57.33	79.80	60.00

Table 2: Evaluation results

	CP Negative	CP Neutral	CP Positive
Key Negative	304	532	113
Key Neutral	154	1458	341
Key Positive	88	595	665

Table 3: Confusion matrix for ClimaPinion on SentiStrength Corpus

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 and Bontcheva, 2015). The evaluation thus investigates to what extent these kind of additions are useful.

The first baseline ARCOMEM (Maynard and Hare, 2015; Maynard et al., 2012)] is an opinion mining tool that was developed in GATE for use in the EU ARCOMEM project[5]. It essentially comprises the core GATE opinion mining tools before the enhancements for ClimaPinion were developed. This acts as a good baseline for ClimaPinion: it is not tuned to the environmental domain and is less sophisticated, but uses the same essential principles.

The second baseline we use (Gindl et al., 2010), which we shall refer to as DIVINE, 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 et al., 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 ClimaPinion, but in contrast, less linguistic analysis is done on the text itself and more reliance is made on the lexicon.

SentiStrength (Thelwall et al., 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 et al., 2012) on this genre (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 easy, and for others to reuse, 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. Note 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.

Results are shown in Table 2. We can see that SentiStrength and ClimaPinion performed best on the datasets which were developed with them in mind, which is unsurprising. ClimaPinion works best on the domain-specific dataset that was manually annotated, rather than the crowdsourced one, which highlights the potential bias of using only a single annotator. In the following sections, we look more closely at the results for each corpus.

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

	SS Negative	SS Neutral	SS Positive
Key Negative	449	326	174
Key Neutral	257	1038	658
Key Positive	33	224	1023

Table 4: Confusion matrix for SentiStrength on SentiStrength Corpus

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

Table 5: Confusion matrix for human annotators on the Earth Hour 2015 corpus

4.1. SentiStrength corpus

Looking at the results on the SentiStrength corpus, the first thing to note is the way the corpus was annotated, and the assumptions made by SentiStrength (Thelwall et al., 2012). In this corpus, posting a URL (without contrary sentiment evidence) is annotated as a positive tweet, 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 in our tools (and manual annotation) 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

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

ClimaPinion's performance. Other instances where we disagree with the gold standard annotations are constructions such as conditionals which demonstrate a type of irrealis mood. For example, in the gold standard SentiStrength corpus, 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.

If we look at the confusion matrices shown in Tables 3 and 4, we also see an interesting distinction. Although SentiStrength has the highest accuracy on this corpus, compared with the other tools, it classifies far fewer tweets than ClimaPinion as neutral. 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.

4.2. Earth Hour 2014 corpus

It is immediately evident that results on the Earth Hour 2014 dataset are much higher for all systems than on the SentiStrength corpus. 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. 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.

4.3. Earth Hour 2015

It is interesting to analyse the differences in performance of the tools between the Earth Hour 2014 corpus, annotated by one person, and the Earth Hour 2015 crowdsourced one, since all other criteria are similar (same domain, same systems, same annotation guidelines etc.). We see that in this dataset, ClimaPinion scores the highest, closely followed by SentiStrength. This differs from the evaluation on the Earth Hour 2014 dataset, where SentiStrength performed much worse comparatively, though with roughly the same actual accuracy score (around 65%). In order to understand why the other 3 systems all perform worse on this dataset than on the Earth Hour 2014 one, we investigated the annotations a little more closely.

The confusion matrix in Table 5 shows how often annotators agreed for each polarity type, and which sentiment classes they confused. We see that there was very little confusion between negative and positive, and not much confusion between negative and neutral, but significant confusion between positive and neutral. This is probably 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). The ClimaPinion tool does not try to annotate such statements as positive, but some of the annotators seemed to find this a difficult distinction to make. In our scenario, the distinctions between negative and positive and between negative and neutral are perhaps the most important to be clear about; our goal is primarily to uncover the level of engagement with the concept of climate change, and therefore the distinction between an overtly positive tweet and a neutral tweet that might still be promoting Earth Hour is actually not so important (or obvious). In some sense, therefore, absolute figures for accuracy in this scenario are less important than considering how well the tools perform on correctly separating negative tweets from neutral and positive ones. This leads to a wider point: the evaluation of sentiment analysis tools always needs to be performed in the context of the application and situation: knowing which tool is most appropriate for the task is more important than having some generally "highest performing" tool for any situation. Almost all sentiment analysis tools work better when adapted to the domain and task, so this stage should not be neglected.

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

Table 6: Confusion matrix for ClimaPinion vs. human annotators on the Earth Hour 2015 corpus

Table 6 shows the confusion matrix for ClimaPinion compared with the gold standard provided by the annotators, for the Earth Hour 2015 corpus. 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 con-

trast, 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 better clarification of annotation guidelines and the positive/neutral distinction.

5. Conclusions

In this paper, we have presented a set of experiments with different sentiment analysis tools and corpora in an attempt to bring to the fore some typical evaluation issues that occur in such contexts. The ClimaPinion tools are released as open-source, along with the Earth Hour 2015 annotated corpus for others to experiment with. It is clear that performance is still not as high on this type of data as on other kinds of text: for example, opinion mining tools typically now score quite highly on product reviews. There are a number of reasons for this, not the least of which is that Twitter is a hard medium to work with, partly because of linguistic pre-processing issues (Maynard, 2014; Derczynski et al., 2015), and partly because tweets offer little semantic context for opinion mining tools (Maynard et al., 2015).

We highlight a number of issues when comparing sentiment analysis tools with each other, and when using manually annotated datasets for the comparison, which can all bias results. First, when evaluating against one's own manually annotated data, there is almost always bias to one's own tools, because annotation is usually done with the criteria in mind that are used for the tool's development. Second, when the domains for training / development and testing are identical, performance will almost always be higher. Comparing other tools which have not been trained on that domain will result in lower performance for them. Even if the tools are designed to work on open-domain text, the nature of the training/testing data may still vary. Related to this is the fact that the manual annotation process is often tailored to a task, dataset or to the annotators responsible, and may therefore cause bias in evaluation results when used by others. Third, when annotated datasets are released for public use, explanations of the annotation process and specifications given to the annotators are often sketchy, if they exist at all, and the user has to guess at some of the decisions made, and/or the reasons for these decisions.

Finally, it is clear from our experiments that the training and testing domains are critical when comparing systems, something which has been acknowledged in tasks such as NER, where systems should be trained on not only the same domain as for testing, but also on recently created data (Augenstein, 2016). This has been rarely addressed for sentiment analysis, due to the lack of available training data and the difficulty of manual annotation.

6. References

I. Augenstein. 2016. *Web Relation Extraction with Distant Supervision*. Ph.D. thesis, University of Sheffield, UK.

Kalina Bontcheva, Ian Roberts, Leon Derczynski, and Dominic Rout. 2014. The GATE Crowdsourcing Plugin: Crowdsourcing Annotated Corpora Made Easy. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Association for Computational Linguistics.

Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphaël Troncy, and Kalina Bontcheva. 2015. Analysis of named entity recognition and linking for tweets. *Information Processing and Management*, 51:32–49.

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

Miriam Fernández, Gregoire Burel, Harith Alani, Lara Schibelsky Godoy Piccolo, Christoph Meili, and Raphael 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*, Copenhagen, Denmark.

Stefan Gindl, Arno Scharl, and Albert Weichselbraun. 2010. Generic high-throughput methods for multilingual sentiment detection. In *Digital Ecosystems and Technologies (DEST), 2010 4th IEEE International Conference on*, pages 239–244. IEEE.

A. Go, R. Bhayani, , and L. Huang. 2009. Twitter sentiment classification using distant supervision. Technical Report CS224N Project Report, Stanford University.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Diana Maynard and Kalina Bontcheva. 2015. Understanding climate change tweets: an open source toolkit for social media. In *Proceedings of EnviroInfo*, Copenhagen, Denmark.

Diana Maynard and Jonathon Hare. 2015. Entity-based opinion mining from text and multimedia. In *Advances in Social Media Analysis*, pages 65–86. Springer.

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

Diana Maynard, Gerhard Gossen, Marco Fisichella, and Adam Funk. 2015. Should I care about your opinion? Detection of opinion interestingness and dynamics in social media. *Journal of Future Internet*.

Diana Maynard. 2014. Challenges in Analysing Social Media. In Adrian Duşa, Dietrich Nelle, Günter Stock, and Gert G. Wagner, editors, *Facing the Future: European Research Infrastructures for the Humanities and Social Sciences*. SCIVERO Verlag, Berlin.

B. Pang and L. Lee. 2008. Opinion mining and sentiment analysis. *Information Retrieval*, 2(1).

Hilke Reckman, Cheyanne Baird, Jean Crawford, Richard Crowell, Linnea Micciulla, Saratendu Sethi, and Fruzsina Veress. 2013. teragram: Rule-based detection of sentiment phrases using sas sentiment analysis. In *Proceedings of SemEval 2013 International Workshop on Semantic Evaluation*, Atlanta, Georgia.

Hassan Saif, Miriam Fernandez, Yulan He, and Harith

Alani. 2013. Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold.

Philip J Stone, Dexter C Dunphy, and Marshall S Smith. 1966. The general inquirer: A computer approach to content analysis.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2012. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173.