How Green is Sentiment Analysis? Environmental Topics in Corpora at the University of Turin

Cristina Bosco[1,4], Muhammad Okky Ibrohim[2], Valerio Basile[1] and Indra Budi[3]

[1]Dipartimento di Informatica - Università degli Studi di Torino, Italy
[2]langing.ai, Indonesia
[3]Faculty of Computer Science - Universitas Indonesia, Indonesia
[4]CRISIS - Centro di Ricerca Interuniversitario sui cambiamenti Socio-ecologici e la transizione alla Sostenibilità, Italy

Abstract

Despite the unanimous recognition of the plight associated with environmental phenomena and the proliferation of the discourse about it, there is still little work on these issues in the field of NLP. This paper provides a report on the activities we are carrying on at the University of Turin in the application of Sentiment Analysis to environmental topics. In pursuit of the goal of developing resources and tools specifically designed for addressing the complexity of the ongoing environmental debate, we are currently focused on exploring the language used for green issues and defining some annotation schemes that can describe them at different granularity.

Keywords

environment, corpora, sentiment analysis

1. Introduction

It has become increasingly common to apply Sentiment Analysis (SA) and text classification to issues with social impact about which people debate. On the one hand, studying a socially impacting phenomenon from such a computational perspective means creating a precise conceptual and linguistic model, thereby achieving a greater understanding of its characteristics, its dynamics, and, not least, how people perceive it. On the other hand, it is a matter of creating tools that can help policymakers and citizens define strategies to address the problems associated with the phenomenon, bearing in mind that the impact of an intervention depends meaningfully on how it is proposed by governments and political parties and accepted by citizens.

Among the issues that have a unique social importance today are certainly those related to the environment in which we live. As far as the emergency related to the environment, at first sight, one cannot but notice that the environmental issues underlie a great complexity. This is due to the mixing of natural and human entities and related interests, such as individuals, public and private organisations on the one side, and climate, animals and plants on the other one. The language used to describe and discuss environmental topics also mirrors this complexity and is featured by a certain degree of specialization.

Modelling this reality can be therefore especially complex but also particularly useful because it ultimately allows us to better understand the relationship between humans and the environment and to be more aware of the sensitivity towards the environment which is hidden in us.

The characteristics of the discourse about the environment can make especially challenging the classification of opinions expressed about it. We may hypothesize that an accurate annotation of data about environmental topics can be helpful in order to achieve reliable results, e.g., in the detection of the polarity or stance in these texts. According to this hypothesis, we are following two major directions: a) to preliminary analyze the linguistic features of the discourse about the environment carried on in different text genres and b) to design specific annotation schemes that take into account the specific features of these texts and to apply them on selected corpora.

The first direction allowed us to better understand the meaning of the wide-spreading discussion about the language used in *green* communication. This was also useful in preparing the ground for the second direction of research, in which we want to model a specific form of communication about *green* issues, namely that realized in social media. Notwithstanding the relevance of the topics we are addressing, in agreement with the results of the systematic survey of the studies about SA applied to the environment [1], it can be observed that currently in this research area there is a gap and we want to fill it out. Only a few projects indeed exist, also for English, in

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
cristina.bosco@unito.it (C. Bosco); muhammadokky.ibrohim@unito.it (M. O. Ibrohim); valerio.basile@unito.it (V. Basile); indra@cs.ui.ac.id (I. Budi)
0000-0002-8857-4484 (C. Bosco); 0000-0002-6943-6553 (M. O. Ibrohim); 0000-0001-8110-6832 (V. Basile); 0000-0002-2107-6552 (I. Budi)

which environmental topics are addressed by applying SA and in which only fairly rough techniques were used.

In this paper, we describe a variety of experiences carried on at the Department of Computer Science of the University of Turin in the development of corpora and tools for SA applied to environmental topics during the last few years.

The paper is organized as follows. The next section briefly surveys previous work related to the application of SA to environmental topics. Section three focuses on the collection of data, while the fourth is about the annotation schemes we adopted. Finally, the last section provides some conclusions and hints about our future works.

2. Background

There is a huge amount of divulgation and communication about environmental issues related in particular to products and services. A 2020 EU Commission study found that more than half of the environmental claims examined in the EU were vague, misleading or unfounded, while 40% were completely unfounded[1]. In section 3.3, we moreover show that it can be difficult for citizens to understand the exact meaning of texts discussing issues related to the environment, making easier to mislead their content.

To explore SA applied to environment topics, researchers have conducted reviews and surveys providing different perspectives. In particular, in [2], a review is conducted to explore the application of SA in the climate change debate. [3] explore the use of SA for analyzing opinions on several smart city issues like climate change, urban policy, energy, and traffic. While [2] explore papers that used various types of data sources (i.e. news articles, social media, etc.), [3] explore only papers that analyze sentiment in social media. However, both [2] and [3] do not provide an in-depth exploration of the NLP techniques (from the creation of dataset to the evaluation of SA models) that researchers used applying SA on natural environment topics, since they only cover a few among the large variety of topics closely related to nature and environment, like food or carbon issues.

3. Exploring *Green* Language

The first step in our investigation consisted of a linguistic analysis of the discourse about the environment and we applied it to documents from public institutions or online journals to inform citizens about these topics. Applying a multilingual perspective we collected texts from an institutional website in Italian and English, and from some Italian journals in which are discussed environmental topics. The first sample of data, described in section 3.1, is the result of a random collection while the second one, described in section 3.2, is collected using keywords about a specific topic related to the environment, i.e. livestock.

3.1. European Environment Agency

The *European Environment Agency*[2] (EEA) is an agency of the European Union that delivers knowledge and data to support Europe's environment and climate goals. Since 1994, EEA and the *European Information Network Environmental training and observation* [3] (Eionet) provides data and information on Europe's climate and environment to citizens and decision-makers European politicians, publishing articles and more extensive reports which address the state of air quality, or a set of inter-connected or systemic issues, such as the mobility system.

We collected Italian and English data from the EEA website and we built two comparable corpora composed of 10 reports each. The Italian corpus (henceforth EEA-Ita) includes 14,612 tokens corresponding to 556 sentences, while the English corpus (henceforth EEA-Eng) is composed of 11,778 tokens corresponding to 562 sentences.

A qualitative analysis based on the lists of frequency, obtained with SketchEngine, shows that the most used terms in both corpora, Italian and English, refer to the theme of sustainable-environmental quality, but with a slight nuance that differentiates the Italian with respect to English. The most frequent terms in the Italian corpus concern especially the sphere of the fight against the conservation of oceans and seas, the sustaining of the Earth's ecosystem and conservation. In the English corpus, instead, we find a higher frequency of terms related to climate change. In both cases, these are not terms of high specialisation, that is, terms that are difficult to understand by the great majority of citizens, but technical terms relating to the field of reference, and therefore not easily traceable in other contexts. For example, in the Italian corpus, we can highlight words such as "siccità" (drought), "effetto serra" (greenhouse effect), "ecosistema" (ecosystem), "inquinamento" (pollution), "suolo" (soil), "microplastiche e nano plastiche" (microplastics and nano plastics), while in the English one "pollution", "climate change", "adaptation", "mitigation", "habitat".

3.2. Livestock Issues

The livestock sector is currently at the center of a heated debate that has focused mainly on intensive farming. Among the several publications in which these issues are

[1] https://quifinanza.it/green/stop-al-greenwashing-in-etichetta-cosa-vuole-fare-lue/699054/

[2] https://www.eea.europa.eu/en

[3] https://www.eionet.europa.eu/

presented and discussed, we selected a sample of texts from online journals, namely mostly from *CREA Futuro* but also from *L'informatore agrario* and *agricultura.it*. Our corpus is composed of 20,854 words (4,386 different lemmas) corresponding to 24,383 tokens, organized into 725 sentences and 21 documents.

CREA Futuro is an initiative of CREA (*Consiglio per la Ricerca in Agricoltura e l'analisi dell'Economia agraria*)[4], the leading Italian research organization dedicated to the agri-food supply chains, supervised by the Ministry of Agriculture, Food Sovereignty and Forests, and organized in 12 research centres. This online publication[5] is aimed at citizens to combine authoritative information, based on scientific evidence. From the CREAfuturo website, we selected a sample composed of 11 documents. The other texts are from the freely accessible web version of two journals, namely *L'informatore agrario*[6] (8 documents) and *agricultura.it*[7] (2 documents).

As expected the frequency lists collected using SkecthEngine show that the words occurring more than 40 times are "produzione" (production), "animali" (animals), "carne" (meat), "acqua" (water), "latte" (milk), "allevamento" (farming), "zootecnia" (livestock), "benessere" (welfare) and "stress.

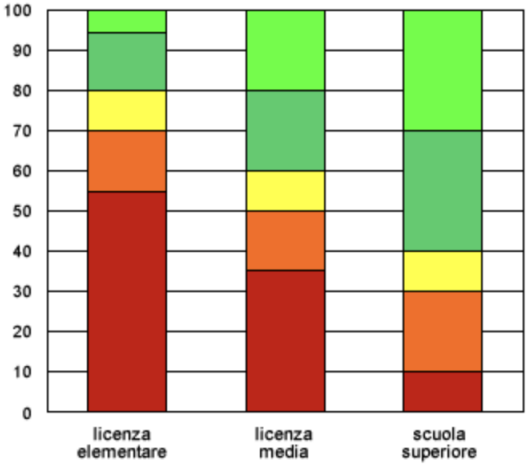

Figure 1: The scales for readability according to the Gulpease index for the three main levels of schooling (primary, secondary and high school): indexes in red for almost unreadable, in orange for very hardly readable, in yellow hardly readable, in dark green easy readable and in light green very easy readable.

3.3. How difficult is to read *green* texts?

All the texts we collected about *green* topics are intended for a general audience, but we want to understand how specialized they are, and thus less or more readable for a citizen. We calculated the readability scores for each of them. Different metrics are used for expressing the readability of different languages and we selected two of the most used ones for the two observed languages.

For Italian texts, we used the Gulpease index[8] whose scales are reported in Figure 1. The Gulpease index has been separately calculated for the 10 reports of the EEA-Ita corpus, showing values that vary from 45 to 53, for the less and the more readable text respectively (see Table 1). This means that the reports are unreadable for readers having primary school diplomas, but hard readable for readers having secondary school diplomas and easily readable for the other ones. According to this index, our texts are on average readable and not particularly specialized with the exception of some terms.

The Gulpease index was calculated also for the 21 documents of the Livestock-Ita corpus showing that are also less readable than the EEA's reports. Considering that the index of the harder-to-read document has a Gulpease index of 28 and the easier an index of 45, they are also featured in a larger variation.

Finally, we used the Flesch–Kincaid index[9] for evaluating the readability of English texts. The values of this index broadly correspond to those of the Gulpease index: values from 100 to 90 are associated with very easy readable texts, from 89 to 80 with easy readable, from 79 to 70 with fairly easy readable, and from 69 to 60 with standard readable. Values below 59 are instead associated with difficult-to-read texts: from 59 to 50 fairly difficult, from 49 to 30 difficult and from 29 to 0 very difficult or almost unreadable without a higher level of schooling.

Corpus	Max G	Min G	Var G
EEA-Italian	53	45	8
lives-Italian	45	28	17
	Max F	**Min F**	**Var F**
EEA-English	46.25	20.24	26.01

Table 1
Indexes of readability: Gulpease index for Italian data (EEA and livestock issues) and Flesh–Kincaid index for English data (EEA).

For English EEA's reports, the Flesch–Kincaid index score varies from 20.24 to 46.25, calculated for the less and the more readable text respectively. This means that

[4] https://www.crea.gov.it/en/home
[5] https://creafuturo.crea.gov.it/
[6] https://www.informatoreagrario.it/
[7] https://www.agricultura.it/
[8] The index can be calculated using the formula provided in [4] and implemented in online calculators, such as https://www.webandmultimedia.it/site/index.php?area=5&subarea=1&formato=scheda&id=36.
[9] This index is described in [5].

the same typology of texts observed for Italian is featured by a higher specialization and meaningfully lower readability. The harder-to-read reports are suitable only for post-graduated people, but also the less difficult ones can be hard to read for undergraduate people.

4. Developing corpora from social media about environmental topics

The observations based on texts published by EEA and in online journals helped us in having a more clear idea of how the language is used for communicating with the citizens and discussing environmental topics. Similar topics are discussed also in social media and we collected data from Twitter in order to build some datasets useful for advancing the application of classification tasks and SA on environmental topics.

Italian data: We collected from Twitter, in a time slot spanning from February 2nd 2022 to March 4th 2022, a total of 8,756 (including some duplicated messages in which more than one of the keywords occurs). They were filtered using the following set of keywords: "Transizione energetica" (energy turnaround), "Agenda 2030", "Crisi climatica" (climate crisis), "Combustibili fossili" (fossil fuel), "Deforestazione" (deforestation), "Greenwashing", "Riscaldamento globale" (global warming), "Impatto ambientale" (environmental impact), "Climate Change", "Green Deal", "Sviluppo sostenibile" (sustainability), "COP26", "Energie rinnovabili" (renewable energy).

English data: we collected from Twitter, within the date range 12 September 2022 until 30 September 2022, a larger amount of data. In collecting this dataset, we used 120 queries from 10 environmental topics including "Environment", "Green", "Sustainability", "Food", "Organism", "Climate Change", "Carbon", "Energy", "Waste", and "Pollution". These 10 environmental topics are obtained from the systematic review conducted by [1], while the queries are obtained from the surveyed papers. We obtained a total of 495,970 tweets, including several duplicated messages, since we use many keywords to collect the data.

4.1. Annotation Schemes for Environmental Topics

We applied three different forms of annotation to our data: one is based on the stance of the user against or in favour of the environmental topics and related politics, one is a fine-grained structured sentiment analysis annotation, while the last one is a sentiment term extraction annotation. The first and second schemes have been applied to the Italian data only, while the last scheme has been applied to the English corpus.

As far as **stance** is concerned, we used the basic scheme based on 3 labels, i.e. Against, Favour, Neutral, also considering Off-topic for the annotation of unclear messages.

In the **fine-grained structured SA** scheme, there are instead two label types that need to be annotated i.e. Spans and Relations. While Span labeling means to identify a set of adjacent or closely connected words, Relation labeling means to identify a relation between two entities annotated as Spans.
Each Span may represent a Holder, an Expression, a Target, or a Topic. A Holder can be a Citizen (an ordinary person/group not affiliated with any official community/organization), a Government (a central or sub-unit government or its stakeholders), a Political Party (a political party or its stakeholders), a Media (a mass media or its stakeholders), a Company (a company or its stakeholders), a Private Foundation (a private foundation or its stakeholders), or an NGO (Non-Governmental Organization). An Expression can be Positive or Negative. The same entities that can be annotated as Holders can be annotated also as Targets. Topics include the general label Environment, but also more specific labels, i.e., the 10 environmental topics we used to collect the English dataset obtained from [1].

Relations are used for labeling the relationship between the Expression and its Holder, Target, or Topic. This allow us to group the Expression and its proper Holder, Target, or Topic, also considering that one tweet can include more Expressions and each Expression may be to be linked to a different Holder, Target and Topic. We also annotate the Coreference as the additional relation label. For the annotation of this fine-grained structured SA annotation, we used the annotation tool provided by Langing Annotate[10]. The example of annotation for this fine-grained scheme can be seen in Figure 2: the text con-

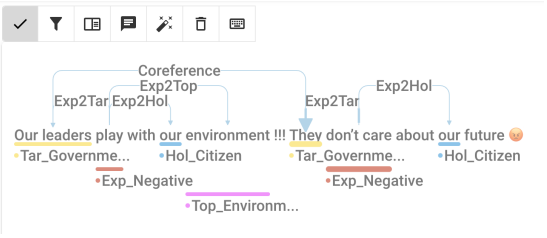

Figure 2: Example of fine-grained structured sentiment analysis

tains two Expressions of negative sentiment. If we wrap each Expression and its Holder, Target, and Topic using a quintuple format (similar to quadruple format used in

[10] https://annotate.langing.ai/

Text	Label
18 gradi a febbraio e rompete i coglioni col riscaldamento globale.. Ne vorrei 30 fissi (18 degrees in February and bust your balls with global warming.. I'd like 30 fixed)	Against
Bottigliette di plastica e collaborazione per ridurre l'impatto ambientale (Plastic bottles and collaboration to reduce environmental impact)	Favour
"#ClimateChange Nel 2021 la crisi climatica è costata 343 miliardi di dollari a livello globale (#ClimateChange In 2021, the climate crisis cost $343 billion globally)	Neutral
Interisti state rosicando così tanto che contribuite alla deforestazione della foresta Amazzonica. #InterMilan (Interisti are so gnawed that you contribute to the deforestation of the Amazon rainforest. #InterMilan)	Off-topic

Table 2
Example of stance annotation.

[6]), i.e. *(Holder, Target, Topic, Expression, Polarity)* we will get two quintuple as follows:

1. *("our", "Our leaders", "environment", "play", negative)*
2. *("our", "They", "", "don't care", negative)*

Notice that in this fine-grained scheme annotation, a Holder, Target, or Topic span should be connected to an Expression span. However, an Expression span can also occur without a Holder, Target, or Topic[11].

Lastly, for **sentiment term extraction** annotation, this scheme is a subset of our fine-grained scheme annotation. Instead of annotating Expression span with its Holder, Target, and Topic, we only annotate the Expression span. Following the guidelines for crowdsourcing datasets conducted by [7], we limit the annotation of English data to Expressions only as a first step, in order to avoid overloading crowdsourcing contributors with a too complex task.

4.2. Annotation of the Italian data

A portion of the Italian data from Twitter, namely 3,254 tweets without duplicates (corresponding to 58,893 words and 1,990 sentences), have been manually annotated for stance, while its annotation with the fine-grained SA scheme is currently ongoing.

4.2.1. Stance annotation

The annotation for this scheme was done using Google Sheets, and some examples of annotation are provided in Table 2.

The agreement occurs in around one-third of the data (2,233 over 3,254), while the disagreement in the other ones (1,021). The higher percentage of disagreement is referred to as the label against, as reported in Table 3. The disagreement has been considered as strong when

	Annotator-1 tweets (%)	Annotator-2 tweets (%)
Against	121 (3.7%)	710 (21.8%)
Favour	1032 (31.7%)	733 (22.5%)
Neutral	1789 (54%)	1691 (52%)
Off-topic	312 (9.6%)	119 (3.7%)

Table 3
Number of labels annotated for each label of the category Stance in the Italian corpus.

Annotator-1 has annotated the message as Against and Annotator-2 as Favour, or vice versa, weak in the other cases. The strong disagreement, occurring in 201 annotated tweets, has been annotated also by a third skilled annotator that solved 168 cases by selecting the label used by the first or that chosen by the second annotator.

4.2.2. Fine-grained structured sentiment analysis annotation

For the annotation of the fine-grained structured SA, we used the same Italian dataset described in Section 4, from which we drew the corpus annotated for stance. In this case, we only selected a portion of the corpus composed of the tweets that contain the keyword "green" (whether a word or subword as in "greenwashing"). Using this filter term, we obtained 1,396 tweets and after dropping the duplicate tweets, we randomly chose 500 tweets to be annotated by two other master's degree students.
For span-level analysis, we analyze the annotation agreement level by calculating the pairwise weighted $F_1 - Score$[12] between annotators using SeqEval library[13]. In this case, $F_1 - Score$ is used to evaluate the span-level agreement because it not only evaluates the entity span agreement but also evaluates the *Beginning, Inside, Outside* (BIO) tagging structure. In this annotation, we obtain

[11]For more examples and details about this fine-grained structured SA annotation see the guidelines: https://github.com/okkyibrohim/environmental-topics-in-corpora/tree/main/annotator_guidelines

[12]We calculate a weighted average of $F_1 - Score$ instead of the macro one since we only annotate 500 tweets for this scheme, making many entities have no enough tweets to be calculated the $F_1 - Score$.

[13]https://github.com/chakki-works/seqeval

a 63.67% of weighted $F_1 - Score$, indicating the annotators have a moderate agreement and can be used for experiments in future works.

To see the sentiment distribution for each annotator, we convert the span-level label to the document-level label into a Negative, Positive, or Neutral, polarity label via majority voting between the Expression label. The distribution of document-level labels between annotators can be seen in Table 4. From Table 4, we see that the sentiment polarity in document-level distribution is quite balanced for Annotator-1. However, in Annotator-2, the Positive polarity has a significant amount more than the other two polarity labels. or this document-level label, we evaluated the agreement score using Cohen's Kappa score and got a score of 0.5718, indicating the document-level label has a moderate agreement and can be used for experiments in future works.

	Annotator-1 tweets (%)	Annotator-2 tweets (%)
Negative	164 (32.8%)	131 (26.2%)
Positive	178 (35.6%)	220 (44.0%)
Neutral	158 (31.6%)	149 (29.8%)

Table 4
Number of labels annotated for each label of the sentiment polarity for document-level in the Italian corpus.

4.3. Annotation of the English data

From the total of 495,970 collected tweets, we randomly select 700 tweets for English sentiment term annotation For this English annotation, we use crowdsourced annotators from Prolific[14] who must have English as their first language and a 100% of approval rate for their previous works in the Prolific platform. Annotators were paid £9/h to perform tasks up to one hour of duration. In this annotation scheme, each data chunk will be annotated by 3 anonymous Prolific workers, which means we have 27 workers in total.

The Fleiss' Kappa score for this annotation, computed at the document level as for Italian, can be seen in Table 5.[15]

Data Chunk	Fleiss' Kappa Score	Kappa Interpretation
1	0.4617	moderate
2	0.5374	moderate
3	0.1673	slight
4	0.4510	moderate
5	0.2778	fair
6	0.4048	moderate
7	0.2538	fair

Table 5
Fleiss' Kappa score for each data chunk for English annotation.

5. Conclusion and future work

This paper presents a report on the activities we are carrying on at the University of Turin in the application of SA to environmental topics. Starting with a linguistic analysis of texts extracted from different genres, we are developing data sets for stance detection, fine-grained structured SA, and sentiment term extraction[16]. Notwithstanding the relevance of these topics, very few applications of textual classification techniques and SA has been developed until now. With our activities, we want to start filling out this gap for Italian and English. Nevertheless this is only a starting point and in future work we will address a more extended domain of texts, for example news and interviews, so as to provide a more reliable barometer of sentiments towards climate topics as found in a general audience.

Acknowledgments

The work of English annotation is funded by PUTI Q1 research grant from Universitas Indonesia with number NKB-394/UN2.RST/HKP.05.00/2022.

Muhammad Okky Ibrohim thanks to FSE REACT-EU for PhD Research Projects funding dedicated to GREEN topics on Ministerial Decree 1061/21.

We thank for their contribution the master's degree students Fabiola Summa, Marco Stella, Martina Gagliardi, Gaia Miele and Maria Comandè.

References

[1] M. O. Ibrohim, C. Bosco, V. Basile, Sentiment analysis for the natural environment: A systematic review, ACM Comput. Surv. (2023). URL: https://doi.org/10.1145/3604605. doi:10.1145/3604605, just Accepted.

[2] M. Stede, R. Patz, The climate change debate and natural language processing, in: Proceedings of the 1st Workshop on NLP for Positive Impact, Association for Computational Linguistics, Online, 2021, pp. 8–18. URL: https://aclanthology.org/2021.nlp4posimpact-1.2. doi:10.18653/v1/2021.nlp4posimpact-1.2.

[3] X. Du, M. Kowalski, A. S. Varde, G. de Melo, R. W. Taylor, Public opinion matters: Mining social media

[14] https://www.prolific.co/

[15] All agreement score interpretation used in this research is obtained from [8]

[16] The dataset and code for agreement evaluation can be seen on this GitHub page: https://github.com/okkyibrohim/environmental-topics-in-corpora

text for environmental management, SIGWEB Newsl. (2019). URL: https://doi.org/10.1145/3352683.3352688. doi:10.1145/3352683.3352688.

[4] P. Lucisano, M. Piemontese, Gulpease: Una formula per la predizione della difficoltà dei testi in lingua italiana, Scuola e città 3 (1988) 110–124.

[5] J. Kincaid, R. Fishburne, R. Rogers, C. B.S., Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel", Research Branch Report (1975) 8–75.

[6] J. Barnes, L. Oberlaender, E. Troiano, A. Kutuzov, J. Buchmann, R. Agerri, L. Øvrelid, E. Velldal, SemEval 2022 task 10: Structured sentiment analysis, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 1280–1295. URL: https://aclanthology.org/2022.semeval-1.180. doi:10.18653/v1/2022.semeval-1.180.

[7] M. Sabou, K. Bontcheva, L. Derczynski, A. Scharl, Corpus annotation through crowdsourcing: Towards best practice guidelines, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 859–866. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/497_Paper.pdf.

[8] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977) 159–174. URL: http://www.jstor.org/stable/2529310.