JOURNAL OF SOCIAL COMPUTING
ISSN 2688-5255 03/06 pp139−157
Volume 3, Number 2, June 2022
DOI: 10.23919/JSC.2022.0003

Measuring Social Solidarity During Crisis: The Role of Design Choices

Steffen Eger*, Dan Liu, and Daniela Grunow

Abstract: Building on our previous work, we assess how social solidarity towards migrants and refugees has changed before and after the onset of the COVID-19 pandemic, by collecting and analyzing a large, novel, and longitudinal dataset of migration-related tweets. To this end, we first annotate above 2000 tweets for (anti-) solidarity expressions towards immigrants, utilizing two annotation approaches (experts vs. crowds). On these annotations, we train a BERT model with multiple data augmentation strategies, which performs close to the human upper bound. We use this high-quality model to automatically label over 240 000 tweets between September 2019 and June 2021. We then assess the automatically labeled data for how statements related to migrant (anti-)solidarity developed over time, before and during the COVID-19 crisis. Our findings show that migrant solidarity became increasingly salient and contested during the early stages of the pandemic but declined in importance since late 2020, with tweet numbers falling slightly below pre-pandemic levels in summer 2021. During the same period, the share of anti-solidarity tweets increased in a sub-sample of COVID-19-related tweets. These findings highlight the importance of long-term observation, pre- and post-crisis comparison, and sampling in research interested in crisis related effects. As one of our main contributions, we outline potential pitfalls of an analysis of social solidarity trends: for example, the ratio of solidarity and anti-solidarity statements depends on the sampling design, i.e., tweet language, Twitter-user accounts' national identification (country known or unknown) and selection of relevant tweets. In our sample, the share of anti-solidarity tweets is higher in native (German) language tweets and among "anonymous" Twitter users writing in German compared to English-language tweets of users located in Germany.

Key words: social solidarity; crises; COVID-19; natural language processing

1 Introduction

Whether people show increased pro- or anti-social behavior in online media during crises has important consequences for social cohesion and political mobilization[1–4]. We use the example of migrant solidarity and anti-solidarity expressed on Twitter to investigate these dynamics. Social solidarity, defined as people's willingness to help others and share own resources beyond immediate rational, individually-, group-, or class-based interests[5] is an important form of pro-social behavior that keeps societies functioning during crises[6]. Conversely, expressions of anti-solidarity towards migrants and refugees are a special form of anti-social behavior. More broadly defined, " (…) anti-social acts are those that demonstrate a lack of feeling and concern for the welfare of others. Indeed, successful social interaction and the smooth running of society can only exist if most people do not behave anti-socially "[6]. Importantly, these definitions consider expressed altruistic orientations towards out-groups, in our case migrants and refugees, or rejection thereof, a basic feature of (anti-)social behavior.

• Steffen Eger is with the Center for Cognitive Interaction Technology (CITEC), Bielefeld University, Bielefeld 33615, Germany. E-mail: steffen.eger@uni-bielefeld.de.
• Dan Liu is with the Computer Science Department, Technical University Darmstadt, Darmstadt 64289, Germany. E-mail: dan-liu@web.de.
• Daniela Grunow is with the Faculty of Social Sciences, Goethe University Frankfurt, Frankfurt 60629, Germany. E-mail: grunow@soz.uni-frankfurt.de.
∗ To whom correspondence should be addressed.

Manuscript received: 2021-09-16; revised: 2022-02-16; accepted: 2022-05-15

Since the onset of the current COVID-19 crisis and especially due to the extended lockdowns during which people were mostly confined to their homes, online media became an even more important means to relate to the outside world, consume news, and express political opinions[7, 8]. Thus, (anti-)social behavior online may have increased and gained relevance for (anti-)social behavior in the real-world[9]. In line with this expectation, current research indicates that both pro- and anti-social behaviors increase during crises and that these trends can be observed in both online and real-world settings[3, 10, 11]. It is at present unclear how persistent such dynamics are and how they relate to one another. Within this field of research, natural language processing (NLP) approaches to investigating pro- and anti-social online behavior from content expressed on text-based social media platforms have great potential to inform social scientists and politicians alike, by complementing existing, mostly survey-based evidence regarding potential threats to social cohesion and vulnerable social groups. To date, research on pro- and/or anti-social online behavior suffers, however, from several shortcomings leading to potentially biased conclusions which this study addresses.

First, observations usually start after the onset of crisis, without considering fluctuation in pro- or anti-social behavior under pre-crisis conditions. It thus remains unclear whether the crisis-event is indeed associated with higher levels of pro- and/or anti-social online behavior targeting a particular social group. Changes recorded after the onset of crises may actually correspond in magnitude to pre-crisis dynamics. Still, the conclusions drawn from this research often suggest an alarming dynamic calling for political intervention[12, 13].

Second, focusing on the current COVID-19 pandemic, research has mainly investigated changes in anti-social online behavior, such as the rise of hate speech, documenting increasing defamation against a wide range of individuals and groups, including immigrants and refugees[14, 15]. Whereas an increase in anti-social online behavior is certainly worrying[16], it remains unclear whether it indeed reflects growing levels of one-sided hostility in society towards social groups or rather heightened issue salience. Heightened issue salience, whether directly related or unrelated to the ongoing pandemic may trigger both pro- and anti-social behavior at the same time, leading to different substantial conclusions and calling for different political responses than a one-sided increase in anti-social behavior alone. The former indicates that resources in society can be mobilized to improve the situation of vulnerable groups, for example, migrants and refugees, even under pandemic conditions whereas the latter suggests that a significant share of the population has shifted to the right while others remain unresponsive and thus hard to mobilize. We argue that dynamics of pro- and anti-social behavior can be interpreted more meaningfully when assessing them jointly and in relation to events that might trigger issue salience.

Third, related to the first two points, causal links between a crisis and the rise of (anti-)social behavior are more often claimed than empirically established. Especially when investigating a long-term crisis such as COVID-19, many other events happen concurrently that may relate to a vulnerable group in focus but not to the crisis itself. Establishing causal claims about pandemic effects on (anti-)social behavior towards social groups is thus complex.

Fourth, findings of pro- and anti-social online behavior can be biased by sampling choices and thus need to be interpreted with care. Researchers (from NLP or data science) primarily interested in training a good machine learning model may make choices that are convenient given the task at hand, for example, restricting the language of tweets to be analyzed to English, but this choice potentially affects the findings. We provide examples of factors that influence findings and may thus bias the conclusions drawn from online discourses.

Fifth, most previous research has looked at changes in (anti-)social online behavior during crises for very short time spans. The duration of heightened issue salience or (anti-)social online behavior and its development over time is important, however, to determine the nature and extent of threats against target groups, social cohesion at large, and the need for state intervention. For example, Ref. [17] showed marked fluctuation of anti-social behavior even over a short time span of five weeks, arguing that peaks may be triggered by political events. Short-term outbursts of (anti-)social online behavior may thus be tied to a single political event, which then likely carry a different meaning than long-term developments of ideological divergence regarding a particular vulnerable social group.

To address these points, (1) we establish a pre-crisis baseline and keep the sampling of social media data constant before and after the onset of crisis. This allows detecting changes in dynamics and ensures that the trends we find during crisis are robust for the sampled subpopulation of data. (2) In doing so, our analysis takes both increases in pro- and anti-social online behavior into account. Our findings thus facilitate distinguishing between societal trends in out-group hostility on one hand and issue salience on the other. (3) As argued above, establishing causal claims about pandemic effects on (anti-)social behavior, whether online or in real-world settings, is problematic and potentially deceptive. We address this by analyzing word occurrence statistics in times of issue salience, rather than simply interpreting peaks or distribution changes as "crisis"-related effects. (4) Our paper shows how susceptible findings regarding rising and/or declining trends of migrant solidarity/anti-solidarity during COVID-19 are to sampling strategy of underlying tweets regarding language, country, and involved keywords, which select different subpopulations of online media users. Even apparently benign decisions such as the language of the users or whether online media posts are selected via hashtags or keywords (in case of Twitter) may influence the results. (5) We provide a long-term view on (anti-)social online behavior, based on daily accurate recording of tweets relating to migrant solidarity. Our data thus enable detecting short- and long-term trends of migrant solidarity during different phases of the crisis. They also allow for distinguishing peaks of heightened issue salience from longer-term trends.

This work builds on our recent ACL conference paper on European solidarity during COVID-19[18]. Departing from the earlier paper, which aimed to analyze different forms of European solidarity, we now focus on migrant solidarity as expressed through social and anti-social online behavior.※ Methodology-wise, our initial conference paper focused on data annotation and developing high-quality text classification models. This extension focuses on the five aspects of design choices outlined above, instead.

Our work is structured as follows. In Section 2, we discuss related work, both from the social sciences and from NLP. In Section 3, we describe the Twitter data that we crawled and annotated, using both experts and crowd-workers. In Section 4, we introduce our classification model, which is popular BERT, together with transfer learning strategies including self-learning, where we leverage unlabeled data via model predictions. In Section 5, we detail our experiments. Sections 2−5 have considerable overlap with our ACL conference publication. In Section 6, we provide a novel large-scale analysis of our newly crawled data, which includes the aspects mentioned above.† We conclude in Section 7.

2 Related Work

Social solidarity in the social sciences. In the social sciences, social solidarity, a key form of pro-social behavior, has always been a topic of intellectual thought and empirical investigation, dating back to seminal thinkers such as Rousseau and Durkheim[5]. Whereas earlier empirical research was mostly confined to survey-based[20−23] or qualitative approaches[24−26], computational social science just started tackling concepts as complex as solidarity as part of NLP approaches[3].

In (computational) social science, several studies investigated the European Migration Crisis and/or the Financial Crisis as displayed in media discourses. These studies focused on (1) differences in perspectives and narratives between mainstream media and Twitter[27], and (2) the coverage and kinds of solidarity addressed in leftist and conservative newspaper media[28, 29], as well as (3) relevant actors in discourses on solidarity[30]. While these studies offer insight into solidarity discourses during crises, they all share a strong focus on mainstream media, which is unlikely to publicly reject solidarity claims[28]. Social media, in contrast, allows its users to perpetuate, challenge, and open new perspectives on mainstream narratives[27]. A first attempt to study solidarity expressed by social media users during crises has been presented by Ref. [3]. They assessed how emojis are used in tweets expressing solidarity relating to two

※Throughout Europe and beyond, migrants have been identified as one of the most vulnerable social groups regarding the epidemiological and economic risks related to COVID-19[19]. In addition, according to the EU-funded sCAN project, which monitors digital communication and hate speech, migrants are often considered responsible for the dissemination of the COVID-19 virus[8].

†Our data and code are available from https://github.com/SteffenEger/socialSolidaritydesign.

crises through hashtag-based manual annotation—ignoring actual content of the tweets—and utilizing an LSTM[31] network for automatic classification. Their approach, while insightful, provides a rather simple operationalization of solidarity, which neglects its contested, consequential, and obligatory aspects vis-à-vis other social groups. The data cover only a few days after the occurrence of different crises (Hurricane Irma 2017 and the 2015 terrorist attacks in Paris), providing little insight regarding the potential to mobilize help in the medium or long run. This long-term aspect is of particular relevance with respect to the population affected in the aftermath of Hurricane Irma.

The current state of social science research on social solidarity, including migrant solidarity, poses a puzzle. On one hand, most survey research paints a rather optimistic view regarding social solidarity in Europe, despite marked cross-national variation[21, 23, 32, 33]. On the other hand, the rise of political polarization[34, 35] suggests that the opinions, orientations, and fears of a potentially growing minority in Europe is underrepresented in this research. People holding extreme opinions have been found to be reluctant to participate in surveys and adopt their survey-responses to social norms (social desirability bias)[36−38]. Research indicates that such minorities may grow in times of crises, with both short-term and long-term effects for public opinion and political trust[34, 39]. Our paper addresses these problems by drawing on large volumes of longitudinal social media data that are more likely to also capture extreme positions on the spectrum of orientations regarding solidarity and its contestation over time[40]. Whereas tweets may not be representative of the full spectrum of orientations regarding social solidarity, they are arguably less prone to social desirability bias than survey data. Moreover, the comparison over time allows us to investigate how (anti-) solidarity statements developed before and during the ongoing pandemic. Thus, the pre-COVID-19 discourse serves as a baseline to which the tweet dynamics during the pandemic can be compared. Even if Twitter users are not representative of society as a whole, our data capture real trends among the Twitter community.

Emotion and sentiment classification in NLP. In NLP, annotating and classifying text (in social media) for sentiment or emotions is a well-established task[41−45]. Importantly, our approach focuses on expressions of (anti-)solidarity: For example, texts containing a positive sentiment towards persons, groups, or organizations which are at their core excluding (e.g., regarding migrants) reflect anti-solidarity and are annotated as such. Our annotations therefore go beyond superficial assessment of sentiment. In fact, the correlation between sentiment labels—e.g., as obtained from Vader[44]—and our annotations in Section 3 is only ~0.2. Specifically, many tweets labeled as solidarity use negatively connoted emotion words.

Computational social science in NLP. Recently, there have been several works in the NLP community addressing social science aspects. For example, Ref. [46] used a model from social psychology to interpret stereotypical language in texts, mapping words in a two-dimensional vector space spanned by the dimensions of "warmth" and "competence". Reference [47] provided a framework to suggest annotation labels to social science students with which to better annotate social media posts for the task of whether the tweets support or refute policy measures for COVID-19. Reference [48] studied the evolution of social biases (antisemitism and anti-communism) over time in German parliamentary proceedings using word embeddings. Reference [49] similarly analyzed gender biases over time in a corpus of millions of digitized books. NLP studies that specifically address the effects of COVID-19 are discussed below.

COVID-19 results on pro- and anti-social online behavior and social solidarity. Empirical research suggests that both pro- and anti-social attitudes and online behaviors increased during the present COVID-19 pandemic. Target groups of anti-social behavior include migrants, women, and members of religious communities, i.e., Jews and Muslim[8, 13, 50]. Whereas these studies use various data and methodologies, there is a small but growing field assessing these trends using NLP methods. Reference [51] documented an increase of anti-social online behavior, i.e., anti-Chinese sentiment, during the early stages of the COVID-19 pandemic, based on English language Twitter data and 4chan's Politically Incorrect board posts (4chan.org). The authors found a shift towards blaming China for the outbreak of the pandemic and an increase in sinophobia during the period covered, November 1 2019 to Mach 22 2020. The findings are corroborated by a more recent study conducted by Ref. [12]. The authors investigated anti-social online behavior based on English language tweets targeting Chinese people during the pandemic, covering the period between January 1 and Mach 31 2020. Since both studies focus on anti-social online behavior targeting China and the Chinese people, the situation of migrants or other vulnerable groups is not addressed. Reference [17] studied change in anti-social behaviors during the pandemic by collecting and analyzing a large-scale dataset of COVID-19 related English language tweets and associated comments between March 17 and April 28 2020 (>40 million COVID-19 related tweets). According to their analysis, China and the Chinese people were among the groups severely targeted by anti-social tweets (in a similar vein[51]), together with other groups known to be discriminated, i.e., Muslims. The study did not mention whether migrants were among the groups affected by discrimination. Instead, politicians and global NGOs were identified as targets of anti-social behavior, suggesting that anti-social behavior on Twitter was triggered by political events, a finding corroborated by in our previous research[18]. Reference [17] used a lexicon-based approach to annotate their dataset, supplemented with an open-source content toxicity analysis API.‡ Whereas their findings provided valuable insights into the short-term dynamics of antisocial behavior, especially high temporal fluctuation, it remains unclear whether anti-social behavior increased overall, relative to pre-COVID-19 times, and how it developed in the long-run.

Reference [14] analyzed German and French language accounts and channels that spread COVID-19 related antisemitic messages from January 2020 to March 2021. The authors reported major increases in antisemitic keyword use over time. German language anti-social behavior was most common on Telegram, whereas French language anti-social behavior was most common on Twitter. Since the study focused on antisemitism, pro- or anti-social behavior towards migrants was not assessed. A common finding among these NLP-based studies is the huge fluctuation of anti-social behavior underlying the general time trend. Reference [17] identified political events as triggers for such peaks.

Reference [52] studied social solidarity during COVID-19 by investigating online donation attitudes among survey participants in Kuwait. In this cross-sectional study, respondents were more inclined to donate money to people affected by the COVID-19 pandemic via online platforms when the information provided was inclusive and well-defined regarding location, type of support, and recipient group, including low-wage migrant workers and their families. A study of pro-migrant protests in Germany showed that pro-migrant activism increased during the pandemic, in spite of assembly bans and restrictions, based on activist groups’ combined online and offline mobilization strategies[53]. This research suggested that the situation of migrants remained salient among the wider public during the pandemic because activists continued to put the issue on the agenda. It is unclear, though, whether pro-migrant online activism may also have produced counter-movements on platforms such as Twitter.

In sum, available studies focus on either pro-social or anti-social attitudes/behavior, making it difficult to assess how these trends relate to each other over time[18]. Some of this research addresses migrant solidarity, but these assessments are rare and based on different data and methods. In addition, with the exception of Ref. [13], studies investigating pro- or anti-social online behavior merely provide a short-term view of specific phases during the pandemic, whereas long-term effects and dynamics remain largely unknown. Finally, most research investigating individual online behavior has focused on English language tweets, disregarding online media-user’s native language and geographical location. Potentially different trends of (anti-)solidarity among different social groups cannot be detected.

3 Data and Annotations

We use the unforeseen onset of the COVID-19 crisis, beginning with the first European lockdown, enacted late February to early March 2020, to analyze how social solidarity statements developed before and during the COVID-19 crisis. The pre-COVID-19 pattern serves as a baseline to which the dynamics since the onset of the pandemic can be compared. In order to keep the baseline solidarity debate comparable before and after the onset of the COVID-19 crisis, we confine our sample to tweets with hashtags predominantly relating to two previous European crises whose effects continue to concern Europe, its member states and citizens: (1) migration and the distribution of refugees among European member states, and (2)

‡In contrast, we will use human annotated data together with BERT, which are overall more promising and reliable.

financial solidarity, i.e., financial support for indebted EU countries. The former solidarity debate predominantly refers to the so-called "Schengen" crisis triggered by large refugee flows in 2015 and the living situation of migrants, the latter mostly relates to the Financial Crisis, followed by the Euro Crisis, and concerns the excessive indebtedness of some EU countries since 2010. Importantly, in the analysis of this extended paper (Section 6), our sole focus will be on migrant solidarity.

3.1 Data

Initial dataset (INIT). We crawled 271 930 tweets between 2019-09-01 and 2020-12-31, written in English or German and geographically restricted to Europe (plus the UK), to obtain setups comparable to the survey-based social science literature on European solidarity. For German tweets, we also allowed a non-specified Twitter account location. We only crawled tweets that contained specific hashtags, to filter for our two topics, i.e., refugee and financial solidarity. We started with an initial list of hashtags (e.g., "#refugeecrisis", "#eurobonds"), which we then expanded via co-occurrence statistics. We manually evaluated 456 co-occurring hashtags with at least 100 occurrences to see if they represented the topics we are interested in. Ultimately, we selected 45 hashtags (see appendix of Ref. [18]) to capture a wide range of the discourse on migration and financial solidarity. Importantly, we keep the hashtag list associated with our 270 000 tweets constant over time.[§]

Extended dataset (EXTENDED). We crawled a second dataset that differs from the first one along two dimensions: (1) the data cover a longer period, from 2019-09-01 to 2021-06-30; (2) the data use a uniform sampling strategy with all the hashtags indicated in INIT (and nothing else), as we noticed later that we had accidentally included non-hashtag based tweets in a subset of the data for INIT. Similarly as for INIT, we selected tweets that were (1) either written in English and had account location given as belonging to the EU (or the UK), or (2) written in German (possibly with unknown geographical information). We collected language, topic, and country information for each tweet. Statistics are shown in Table 1. We notice that the majority of tweets are in German and the overwhelming majority (~95%) relates to refugees. Only three countries are of noticeable size: Germany, the UK, and UNK (= unknown), i.e., where no account location is specified by the user; see also Fig. 1, which shows the number of tweets per month of six biggest countries (in terms of number of tweets) and UNK in our dataset.

Table 1 Aspects, values, sizes, and relative fraction of total number of tweets in EXTENDED.

Aspect	Value	Size	Fraction (%)
Language	de	149 107	61.2
	en	94 705	38.8
Topic	Refugee	230 315	94.5
	Financial crisis	13 312	5.5
	Both	185	0.1
Country	Germany	86 407	35.4
	UNK	68 183	28.0
	UK	44 800	18.4
	France	6871	2.8
	Belgium	6629	2.7
	Austria	6482	2.7
	Greece	4158	1.7
	⋮	⋮	⋮
Prediction	`Solidarity`	147 275	60.4
	`Anti-solidarity`	52 191	21.4
	`Other`	44 346	18.2
Total number of tweets	—	243 812	100.0

Note: de stands for German and en stands for English.

We notice that our extended dataset is smaller in size, covering 243 812 tweets, despite stretching over a longer time period (as it only includes hashtag based tweets).

3.2 Annotation

Expert annotations. After crawling and preparing the data, we set up guidelines for annotating tweets. Overall, we set four categories to annotate, with solidarity and anti-solidarity being the most important ones. A tweet indicating support for people in need, the willingness and/or gratitude towards others to share resources and/or help them is considered expressing `solidarity`. The same applies to tweets criticizing the EU in terms of not doing enough to share resources and/or help socially vulnerable groups as well as advocating for the EU as a solidarity union. A tweet is considered to be expressing `anti-solidarity` statements if the above-mentioned criteria are reversed, and/or, the tweet

[§] We follow a purposeful sampling frame, but this necessarily introduces a bias in our data. We will discuss more about this in Section 6.

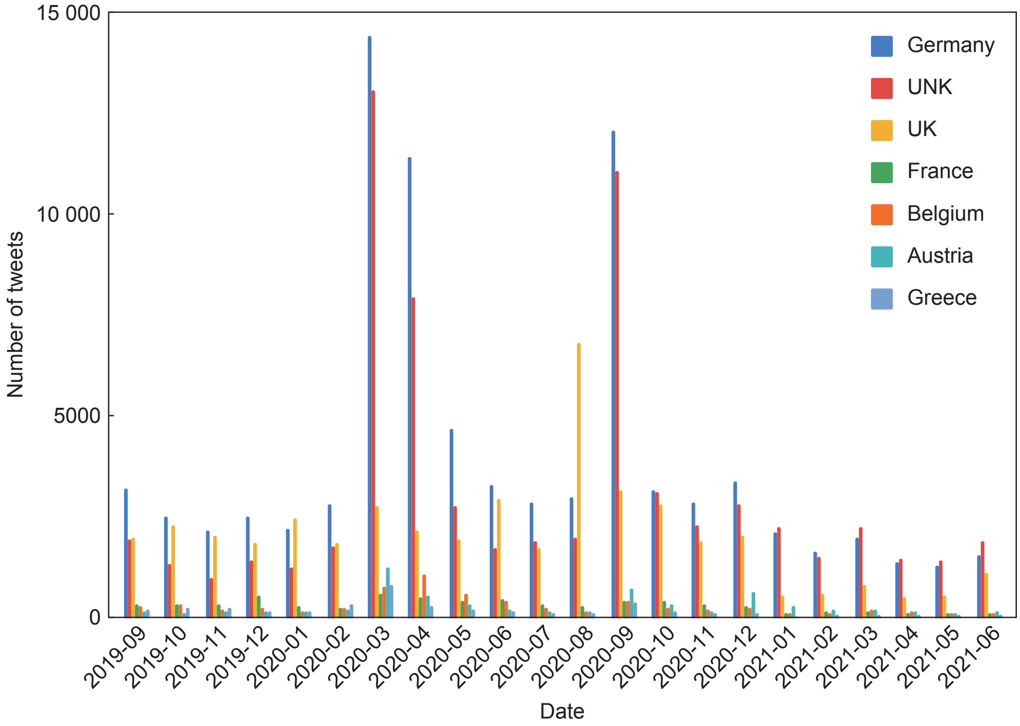


Fig. 1 Number of tweets per month of six biggest countries (in terms of number of tweets) and UNK in our dataset.

contains tendencies of nationalism or advocates for closed borders. Not all tweets fit into these classes, thus we introduce two additional categories: `ambivalent` and `not applicable`. While the ambivalent category refers to tweets that could be interpreted as both expressing solidarity and anti-solidarity statements, the second category is reserved for tweets that do not contain the topic of (anti-)solidarity at all or refer to topics that are not concerned with discourses on refugee or financial solidarity. Table 2 contains example tweets for all categories. Full guidelines for the annotation of tweets are given in the appendix of Ref. [18].

Our annotators included four university students majoring in computer science, one computer science faculty member, as well as two social science experts (one PhD student and one professor). We started the training of seven annotators with a small dataset that they annotated independently and refined the guidelines during the annotation process.

Table 2 Paraphrased (and translated) sample of annotated tweets in our dataset, together with labels.

Tweet	Class label
Children caught up in the Moria camp fire face unimaginable horrors #SafePassage #RefugeesWelcome	`solidarity`
Most people supporting #RefugeesWelcome are racists or psychopaths	`anti-solidarity`
Does this rule apply to every UK citizen as well as every #AsylumSeeker?	`ambivalent`
Let’s make #VaccinesWork for everyone #LeaveNoOneBehind	`not applicable`

While the kappa value was low in the first stages, we managed to raise the inter-annotator reliability over time through discussions (of unclear cases) with the social science experts and extension of the guidelines (e.g., how to handle links in tweets). We also introduced a gold-standard for annotations which served as orientation. This was determined by majority voting and discussions among the annotators. For cases where a decision on the gold-standard label could not be reached, a social science expert decided on the gold-standard label; some hard cases were left undecided (not included in the dataset).

On average across multiple stages of annotation, our kappa agreement is 0.64 for four and 0.69 for three classes (collapsing `ambivalent` and `not applicable` into an `other` class),[¶] while the macro F1-score is 69% for four and 78.5% for three classes. However, in the final stages, the agreement is considerably higher: above 80% macro-F1 for four and between 85.4% and 89.7% macro-F1 for three classes. A summary of agreements is shown in Table 3.

Crowd annotations. We also conducted a “crowd

[¶] We refer to this third class both as `other` and `rest` in the remainder.

Table 3 Agreement levels for expert annotators and crowd-workers. Agreements are for 4/3 classes. Within-1 is the agreement of crowd workers on the new annotations; Within-2 is the agreement of crowd workers on the instances of the gold standard.

Annotator	Aggregation	Macro-F1 (%)	Kappa
Expert	Average	69.2/78.5	0.64/0.69
	Final stages	81.4/87.5	0.78/0.81
Crowd	Wixthin-1	54.9/67.7	0.43/0.49
	Within-2	54.2/68.9	0.45/0.54
	With experts	59.2/77.6	0.49/0.64

experiment” with students in an introductory course to NLP. We provided students with the guidelines and 100 expert annotated tweets as illustrations. We trained crowd annotators in three iterations. (1) They were assigned reading the guidelines and looking at 30 random expert annotations. Then they were asked to annotate 20 tweets themselves and self-report their kappa agreement with the experts (we provided the labels separately so that they could further use the 20 tweets to understand the annotation task). (2) We repeated this with another 30 tweets for annotator training and 20 tweets for annotator testing. (3) They received 30 expert-annotated tweets for which we did not give them access to expert labels, and 30 entirely novel tweets, that had not been annotated before. These 60 final tweets were presented in random order to each student. 50% of the 30 novel tweets were taken from before September 2020 and the other 50% were taken from after September 2020.

125 students participated in the annotation task. The annotation experiment was part of a bonus the students could achieve for the course (counted 12.5% of the overall bonus for the class). Each novel tweet was annotated by up to 3 students (2.7 on average). To obtain a unique label for each crowd-annotated tweet, we used the following simple strategy: we chose either the majority label among the three annotators or the annotation of the most reliable annotator in case there was no unique majority label. The annotator that had the highest agreement with the expert annotators was taken as most reliable annotator.

The distribution of kappa agreements of students with the experts is shown in Fig. 2; summarizing statistics are displayed in Table 3. The majority of students have a kappa agreement with the gold-standard of between 0.6−0.7 when three classes are taken into account and around 0.5 for four classes.

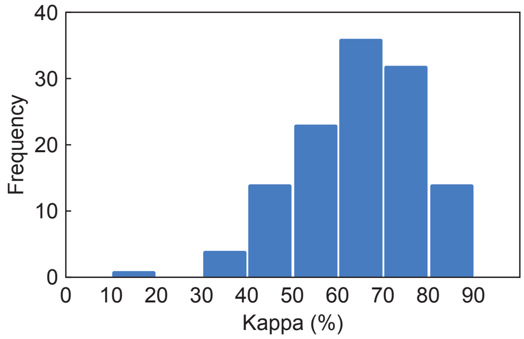


Fig. 2 Distribution of kappa agreements of crowd workers with expert annotated gold standard, 3 classes.

Table 4 Number of annotated tweets (after geofiltering) for the four classes `solidarity` (S), `anti-solidarity` (A), `ambivalent` (AMB), and `not applicable` (NA).

Annotator	S	A	AMB	NA	Total
Expert	386	246	113	174	919
Crowd	768	209	186	217	1380

In Table 4, we further show statistics on our annotated datasets: we have 2299 annotated tweets in total, about 60% of which have been annotated by crowd-workers. About 50% of all tweets are annotated as `solidarity`, 20% as `anti-solidarity`, and 30% as either `not-applicable` or `ambivalent`. In our annotations, 1196 tweets are English and 1103 are German.[1] Finally, we note that the distribution of labels for expert and crowd annotations are different, i.e., the crowd annotations cover more solidarity tweets. The reason is twofold: (1) for the experts, we oversampled hashtags that we believed to be associated more often with anti-solidarity tweets as the initial annotations indicated that these would be in the minority, which we feared to be problematic for the automatic classifiers. (2) The time periods in which the tweets for the experts and crowd annotators fall differ.

4 Method

We use multilingual BERT[54] / XLM-R[55] to classify our tweets in a 3-way classification problem (`solidarity`, `anti-solidarity`, and `other`), not differentiating between the classes `ambivalent` and `non-applicable` since our main focus is on the analysis of changes in (anti-)solidarity. We use the baseline multilingual BERT model: bert-base-multilingual-cased and the base XLM-R model: xlm-roberta-base. We implemented several data augmentation/transfer learning techniques to improve

[1] In our automatically labeled data, the majority of tweets are German. This may be due to our more aggressive geofiltering of English tweets.

model performance:

- **Oversampling of minority classes**: We randomly duplicate (expert and crowd annotated) tweets from minority classes until all classes have the same number of tweets as the majority class `solidarity`.
- **Back-translation** : We use the Google Translate API to translate English tweets into a pivot language (we used German), and pivot language tweets back into English (for expert and crowd-annotated tweets).
- **Fine-tuning**: We fine-tune MBERT / XLM-R with masked language model and next sentence prediction tasks on domain-specific data, i.e., our crawled unlabeled tweets.
- **Auto-labeled data**: As a form of self-learning[56], we train 9 different models (including oversampling, back-translation, etc.) on the expert and crowd-annotated data, then apply them to our full dataset (of 270 000 tweets). We only retain tweets where 7 of 9 models agree and select 35 000 such tweets for each label (`solidarity`, `anti-solidarity`, and `other`) into an augmented training set, thus increasing training data by 105 000 auto-labeled tweets.

We also experimented with re-mapping multilingual BERT and XLM-R[57–59] as they have not seen parallel data during training, but found only minor effects in initial experiments.

5 Experiment

In Section 5.1, we describe our experimental setup. In Section 5.2, we show the classification results of our baseline models on the annotated data and the effects of our various data augmentation and transfer learning strategies. In Section 5.3, we analyze performance of our best-performing models.

5.1 Experimental setup

To examine the effects of various factors, we design several experimental conditions. These involve (1) using only hashtags for classification, ignoring the actual tweet text, (2) using only text, without the hashtags, (3) combining expert and crowd annotations for training, (4) examining the augmentation and transfer learning strategies.

All models are evaluated on randomly sampled test and dev sets of size 170 each. Both dev and test set are taken from the expert annotations. We use the dev set for early stopping. To make sure our results are not an artefact of unlucky choices of test and dev sets, we report averages of 3 random splits where test and dev set contain 170 instances in each case.

We report the macro-F1 score to evaluate the performance of different models. Hyperparameters of our models can be found in our github.

5.2 Result

The main results are reported in Table 5.[**] Using only hashtags for expert annotated data yields a macro-F1 score of just slightly above 50% for MBERT and XLM-R. Including the full texts improves this by over 10 points. Adding crowd-annotations yields another small boost of up to 2 points. Removing hashtags in this situation decreases the performance between 2 and 5 points. This means that the hashtags indeed contain import information, but the texts are more important than the hashtags: with hashtags only, we observe macro-F1 scores of around 50% on the test set, whereas with text only the performance is substantially higher, around 60%. While using " hashtags only" means less data since not all of our tweets have hashtags, the performance with only hashtags on the test sets stays around 50% both with 574 and more than 1500 tweets for training.

Next, we analyze the data augmentation and transfer learning techniques. Oversampling, backtranslation, and pretraining are all similarly effective, and improve the scores on the test set by about 1−4 points. Including auto-labeled data drastically increases the train set, from below 2000 instances to over 100 000. Even though these instances are self-labeled, performance increases by 10−14 points to 75%−78% macro-F1 on the test set. Combining all strategies yields scores of up to 80%.

To sum up, we note: (1) adding crowd annotated data improves the quality of the model, despite the crowd annotated data having a different label distribution; (2) including text is important for classification as the classification with " hashtags only" performs considerably worse; (3) data augmentation (especially self-labeling) and transfer learning strategies have a further clearly positive effect.

5.3 Model analysis

Table 6 shows selected misclassifications for a high-

[**] Those results are different from our ACL conference publication, as we had accidentally kept the test set constant in the conference publication.

Table 5 Macro-F1 scores for different conditions. Entries with ± give averages and standard deviations over 3 different runs with different test and dev sets. "E" stands for experts, "C" for crowds. "ALL" refers to all data augmentation and transfer learning techniques.

Condition	Train size	MBERT		XLM-R	
		Dev (%)	Test (%)	Dev (%)	Test (%)
E, Hashtag only	574	59.5±2.5	52.2±3.0	54.0±5.1	51.3±6.0
E	579	66.8±1.8	64.2±1.9	69.9±1.9	62.5±3.7
E+C	1959	65.6±2.3	65.7±1.7	70.5±2.8	64.1±5.8
E+C, No hashtags	1956	63.3±1.1	60.9±3.8	64.2±2.4	62.0±3.3
E+C, Hashtag only	1529	59.5±2.5	52.2±3.0	53.6±2.4	47.3±6.2
E+C+Oversample	3040	68.1±2.4	66.5±1.3	71.0±4.1	67.3±1.2
E+C+Backtranslation	3854	70.6±3.8	66.9±1.0	73.2±1.7	68.0±2.5
E+C+Pretraining	1959	72.5±4.0	66.6±1.8	72.1±3.3	68.7±1.5
E+C+Auto labeling	106 959	78.2±1.5	75.3±5.0	83.1±1.0	78.2±3.6
E+C+All	109 935	81.0±0.6	77.1±0.3	81.1±2.1	80.9±0.3

Table 6 Selected misclassifications of best performing ensemble model. We consider the bottom tweet misclassified in the expert annotated data (correct would be `solidarity`). Tweets are paraphrased and/or translated. "O" stands for "Other".

Text	Gold	Prediction
(1) You can drink a toast with the AFD misanthropists #seenotrettung #NieMehrCDU	S	A
(2) Why is an open discussion about #Remigration (not) yet possible?	O	A
(3) Raped and Beaten, Lesbian #AsylumSeeker Faces #Deportation	A	O

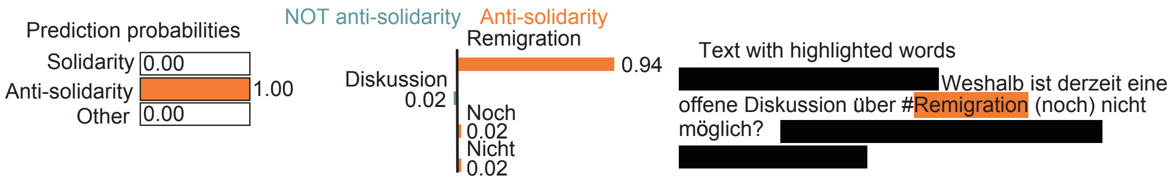
performing model with performance above 80% macro-F1. This reveals that the models sometimes leverage superficial lexical cues (e.g., the German political party "AfD" is typically associated with anti-solidarity towards refugees), including hashtags ("Remigration"); see Fig. 3, where we use LIME[60] to highlight words the model pays attention to. To further gain insight into the misclassifications, we have one social science expert reannotate model misclassifications. From 25 errors that this model made in the test set of 170 instances, the expert thinks that 12 times the gold standard is correct, 7 times the model prediction is correct, and in further 6 cases neither the model nor the gold standard are correct. This hints at some level of errors in our annotated data; it further supports the conclusion that our model is not far from the human upper bound.

6 Trend Analysis

We used our best performing model to automatically label all our 243 000 tweets between September 2019 and June 2021. Figures 4 and 5 show the frequency curves for , `solidarity`, `anti-solidarity`, and `other` tweets over time in our sample on a daily and monthly basis, respectively. Figures 4 and 5 also give the ratio

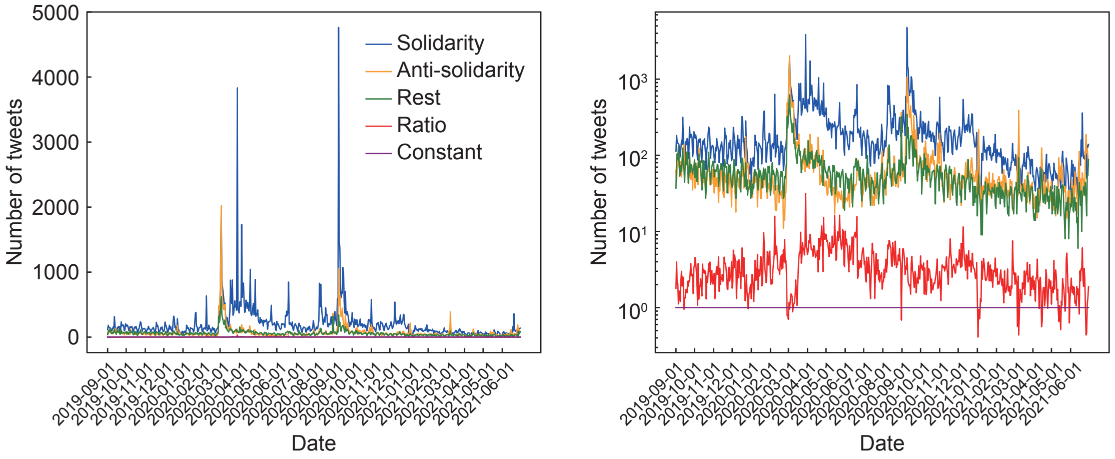
$$\text{S/A} := \frac{\#\text{Solidarity tweets}}{\#\text{Anti-solidarity tweets}},$$

which shows the frequency of solidarity tweets relative to anti-solidarity tweets. Values above (below) one indicate that more (less) solidarity than anti-solidarity



[redacted] Weshalb ist derzeit eine offene Diskussion über #Remigration (noch) nicht möglich? [redacted]

Fig. 3 Picture on the left shows the prediction probabilities for the three classes, the picture in the middle shows features (word) reponsible for the model classification (according to LIME), along with their importance score. The picture on the right shows the original tweet. Our best-performing model predicts `anti-solidarity` for the current example because of the hashtag #Remigration (according to LIME). The tweet, also given as translation in Table 6 (2), is overall classified as `other` in the gold standard, as it may be considered as expressing no determinate stance. Here, we hide identity revealing information in the tweet, but our classifier sees it.



(a) Original scale

(b) Logscale

Fig. 4 Trend curves for `solidarity`, `anti-solidarity`, and `other` over time (daily). Only refugee related tweets.

statements are tweeted.

We note that Figs. 4 and 5 only refer to refugee-related tweets, as these constitute the vast majority in our data and we remove all other tweets in the remainder of our analysis, unless indicated otherwise.

INIT vs. EXTENDED. We first compare the two data distributions (INIT and EXTENDED) that we have crawled. A comparison, for the same underlying time frames, is shown in Table 7. To make the comparison fair, we keep the non-refugee related tweets in EXTENDED here. We see that the Spearman/Pearson correlation between the two time series is very high (0.95−0.99) in the period from September 2019 to August 2020. Small variations may be a result of different crawlers (which find different tweets) and from the non-deterministic nature of the underlying neural networks that we used (as we retrained the networks). For the complete time period from September 2019 to December 2020, the correlations are lower and especially the correlation of the `anti-solidarity` and `other` trend curves is weaker (0.79−0.89 Spearman and Pearson). The reason is the following.

Retrospectively, we discovered that we have (accidentally) sampled tweets that do not contain hashtags in the time period from September 2020 to December 2020 in the INIT dataset. More specifically, we sampled tweets without our hashtags described in the appendix of Ref. [18] and in Section 3, but that still contain corresponding keywords (such as “refugees”). Interestingly, the distribution of solidarity in tweets with and without hashtags is very different, as shown in Table 8. In particular, tweets without hashtags are much more likely to display `anti-solidarity` and much less likely to show `solidarity`. A reason may be that our hashtags have a “solidarity bias”. Another related reason may be that, due to the general setup of social media platforms like Twitter, being supportive and non-offensive is better accepted[61], which is why hashtags with negative association may be less common.

This finding means that design choices such as how to sample tweets may considerably change the observed outcomes and utmost care must be taken to keep such factors constant over time, see our Point 4 in the introduction.

Peaks of `solidarity` and `anti-solidarity` over time. Figure 4 displays several short-term increases in (anti-)solidarity statements in our window of observation. Introspection (see Fig. 6) shows that these peaks have been immediate responses to drastic politically relevant events, which were also prominently covered by mainstream media, e.g., natural disasters, fires, and policy changes. We illustrate this in the following.

On March 11th 2020, the World Health Organization (WHO) declared the COVID-19 outbreak a global pandemic. Shortly before and after, European countries started to take a variety of countermeasures, including stay-at-home orders for the general population, private gathering restrictions, and the closure of educational and childcare institutions[62]. With the onset of these interventions, both solidarity and anti-solidarity statements relating to refugees increased dramatically. At its peak at the beginning of March, anti-solidarity statements outnumbered solidarity statements (we

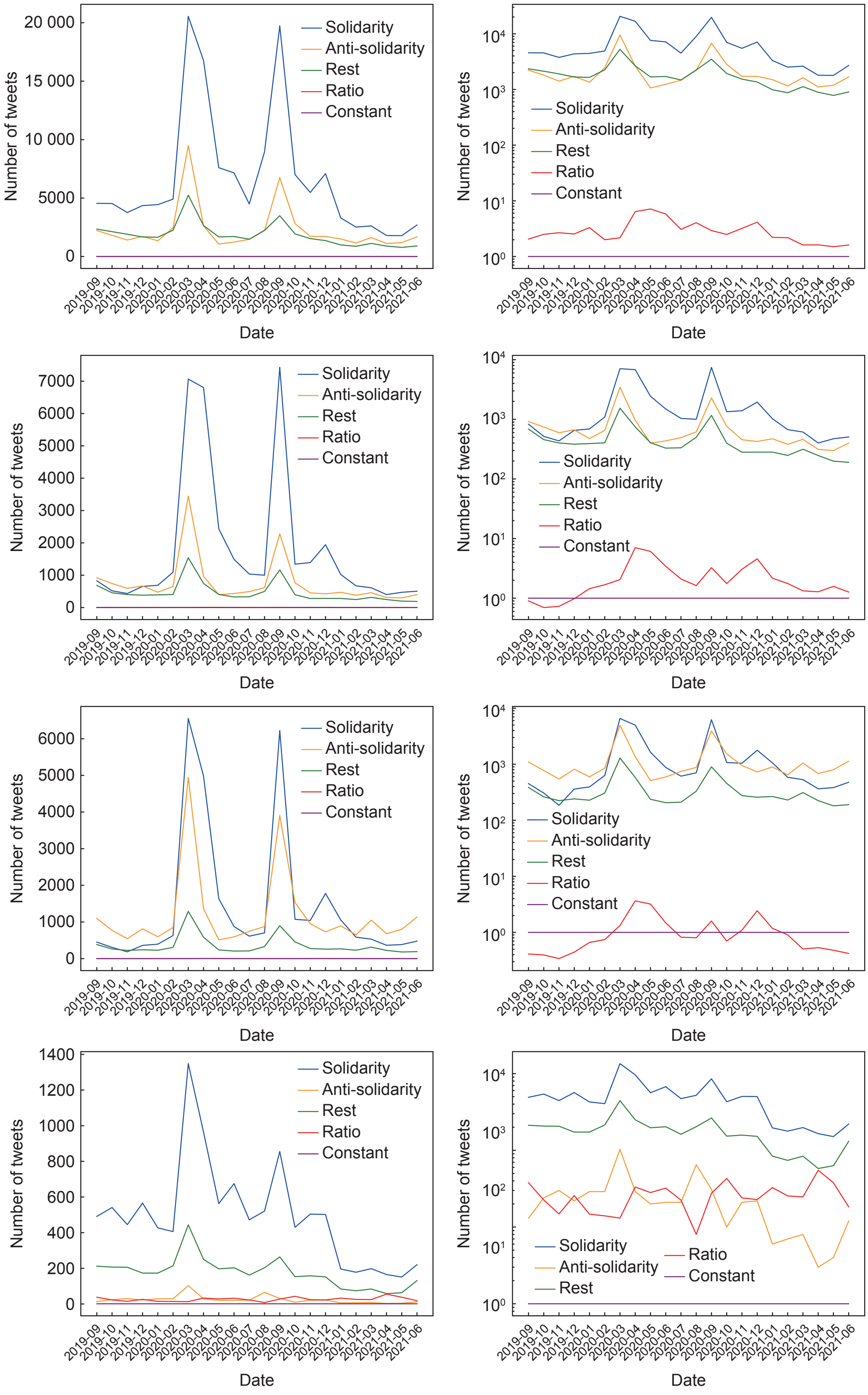


Fig. 5 Trend curves for `solidarity`, `anti-solidarity`, and `other` over time (for refugee related discourse). Monthly aggregates. Left: Original scale, right: Logscale. Row 1: All sampled tweets. Row 2: Country=Germany, lang=de. Row 3: Country=UNK, lang=de. Row 4: Country=Germany, lang=en.

Table 7 Comparison between `solidarity`, `anti-solidarity`, and `other` time series from the two different datasets, INIT and EXTENDED. Spearman / Pearson correlation.

Class label	2019-09-01 to 2020-12-31 (487 days)	2019-09-01 to 2020-08-31 (365 days)
`Solidarity`	0.99/0.99	0.98/0.99
`Anti-solidarity`	0.84/0.89	0.95/0.99
`Other`	0.78/0.87	0.96/0.98

Table 8 Distribution (in %) of `solidarity`, `anti-solidarity`, and `other` in tweets with and without hashtags in time period from September 2020 to December 2020. Average and standard deviation for each of the four months.

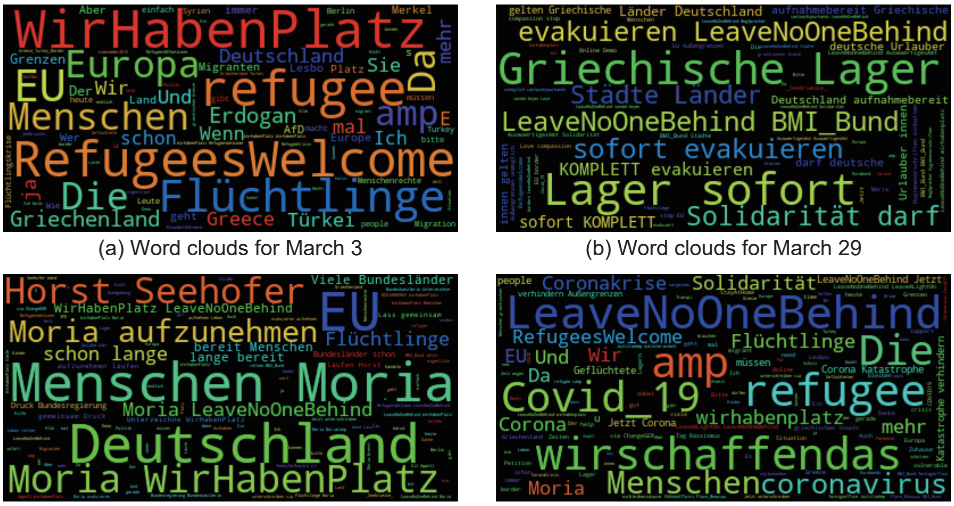
Class label	With hashtags	Without hashtags
`Solidarity`	0.56±0.03	0.21±0.02
`Anti-solidarity`	0.30±0.03	0.53±0.05
`Other`	0.14±0.02	0.26±0.03

recorded below 2000 solidarity tweets vs. above 2000 anti-solidarity tweets on March 3rd). Whereas it may seem straight-forward to conclude that these peaks are related to the lockdown interventions, introspection of the data shows that the latter peak of solidarity and anti-solidarity statements was a reaction to Turkey president Erdogan's declaration to open borders for refugees as a consequence of disputes with the European Union, over how to handle the high inflow of Syrian refugees.

The dominance of solidarity statements was soon reestablished, especially on March 29, when the EU announced funding for new refugee camps in Greece, which led to an outburst of solidarity in our data. Over the following months, anti-solidarity statements decreased again to pre-COVID-19 levels, whereas solidarity statements remained comparatively high, with several peaks between March and September 2020.

Solidarity and anti-solidarity statements shot up again early September 2020, with an unprecedented climax on September 9th. Introspection of our data shows that the trigger for this was the precarious situation of refugees after a fire destroyed the Mória Refugee Camp on the Greek island of Lesbos on the night of September 8th. Human Rights Watch had compared the camp to an open-air prison in which refugees lived under inhumane conditions, and the disaster spurred debates about the responsibilities of EU countries towards refugees and the countries hosting refugee hot spots (i.e., Greece and Italy).

Figure 5 offers a monthly perspective on the same data. In the top part (all tweets), we observe a relative increase in solidarity from March 2020 to roughly December 2020. After this period, the number of tweets on refugee related topics decreases and the relative proportion of `solidarity` and `anti-solidarity` tweets is at levels comparable to before March 2020. There is also often a simultaneous



(a) Word clouds for March 3

(b) Word clouds for March 29

(c) Word clouds for September 9

(d) Word clouds for March 20

Fig. 6 Word clouds for March 3, March 29, September 9, and March 20. The discussion on March 3 was heavily influenced by Turkey president Erdogan's decision to open borders for refugees to the EU. March 29 marks a wave of solidarity with refugees after the EU announced the construction of new refugee camps in Greece, considered inhumane among a larger public. September 9 marks the burning of camp Moria in Greece. March 20 was one of the days with highest saliency of COVID in the refugee discourse (as confirmed by the word cloud). Interestingly, the discourse on such days of heated debates is dominated by German speaking tweets, possibly indicating a higher saliency of these topics in Germany.

increase of solidarity and anti-solidarity tweets over the months, which indicates a polarized discourse and heightened issue salience. We note that solidarity outweighs anti-solidarity here consistently for each month.

In sum, our findings here address Points 1− 3 and 5 laid out in the introduction: We frequently observe a simultaneous increase in both solidarity and anti-solidarity over time during crisis; these fluctuations are not always related to the pandemic, illustrating the problem of causal interpretation. Finally, given our pre-crisis baseline, we clearly observe an initial increase in (anti-)solidarity in our data which appears to subside over longer time spans.

The influence of language and country. As discussed, Fig. 5 presents a monthly perspective on the data. Row 1 presents results over time for all of our data (English and German language tweets from all over Europe, including the UK). We then subsample tweets: we select German language tweets, separating tweets from users who indicate Germany as their location (row 2) and users who do not provide any information on their location (row 3). We refer to the later group of users as anonymous.[††] Finally, we look at English language tweets posted by users who indicate Germany as their location (row 4).

We observe that users from Germany that tweet in German show relatively more anti-solidarity than our full sample. This is indicated by the ratio line being closer to the constant line in row 2 compared to row 3, on the right-hand side of Fig. 5. When tweets are in German but have no location information, then tweets show even more anti-solidarity, but the trend curves over the time period are very similar between the latter two conditions. In contrast, the sample of users who tweet in English and whose country location is given as Germany show considerably more solidarity than either of the other two groups. We speculate that users who tweet in English are either foreigners, have a better education, or a lower national identity, among others. Since the majority of tweets outside Germany are also in English, this may explain why the overall trend shows more solidarity than in the German sample. We speculate that if we had more native language tweets from other countries, these could likewise show more anti-solidarity, on average. A similar finding is reported in Ref. [63], i.e., that language and (political) identity may interact in social media expressions.

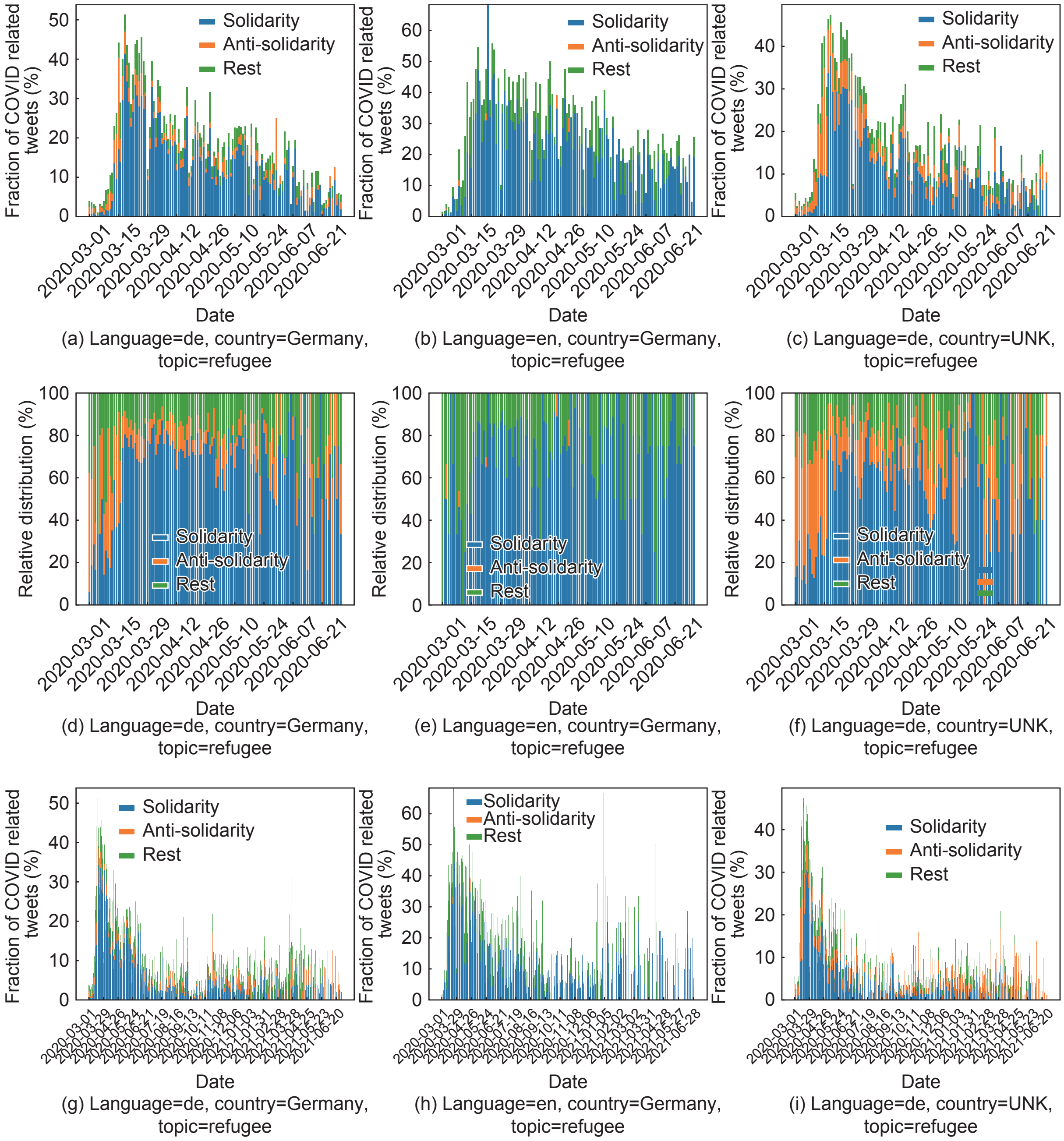
This finding speaks again to Point 4 from the introduction, namely, that sampling decisions may lead to drastically different conclusions. A way to deal with sampling bias includes using representative survey data concerning social solidarity as "ground truth", to double-check the validity of measurements obtained from online data. However, as outlined in Section 2, survey data on social solidarity are biased as well, i.e., by selective response to sensitive items and socially desirable responses. Another option is to disclose and acknowledge potential biases or to measure stability of results over different sampling choices.

Refugee-related tweets that contain COVID keywords. Figure 7 looks at refugee related tweets that contain relevant COVID keywords (COVID, Corona, and virus, in both upper and lower case). We observe that the COVID pandemic "infects" the refugee discourse in late February / early March 2020. There is a sudden increase until more than 40% of all refugee-related Tweets contain COVID-related terms by mid/end of March 2020 (row 1 and row 3). This "shock" subsides rather quickly—by end of June 2020, less than 15% of all refugee-related tweets include COVID-related terms and the frequency then roughly stabilizes at rather low levels. When looking at the relative distribution of solidarity, anti-solidarity, and other in those refugee-related tweets that contain COVID terms (row 2), we observe a similar pattern as in Fig. 5: Users whose location is unknown and speak German (right-most figures) show most anti-solidarity and users whose location is given as Germany and who speak English show most solidarity (middle figures).

It is noteworthy, however, that all three subpopulations considered show more solidarity in the heated initial phase of the pandemic (mid-March to roughly end of April 2020), but two subpopulations (both involving German language tweets) show more anti-solidarity the longer the pandemic lasts (with low levels of statistical support however).

This analysis also shows that the direct association between refugee related tweets and the COVID-19 crisis quickly diminishes over time (though it remains permanent in our data at low levels), weakening causal interpretations. This speaks to our initially raised

[††] This is account location, which users can set arbitrarily or also have the choice to omit.



(a) Language=de, country=Germany, topic=refugee

(b) Language=en, country=Germany, topic=refugee

(c) Language=de, country=UNK, topic=refugee

(d) Language=de, country=Germany, topic=refugee

(e) Language=en, country=Germany, topic=refugee

(f) Language=de, country=UNK, topic=refugee

(g) Language=de, country=Germany, topic=refugee

(h) Language=en, country=Germany, topic=refugee

(i) Language=de, country=UNK, topic=refugee

Fig. 7 Fraction of tweets that contain COVID related keywords among all "refugee" related tweets over time. Top: 2020-03-01 to 2020-06-30, fraction of `solidarity`, `anti-solidarity`, and `other` among all tweets. Bottom: 2020-03-01 to 2021-06-30, fraction of `solidarity`, `anti-solidarity`, and `other` among all tweets. Middle: 2020-03-01 to 2020-06-30, relative distribution of `solidarity`, `anti-solidarity`, and `other`.

Points 2, 3, and 5 in the introduction.

7 Conclusion

In this paper, we contribute the first large-scale human and automatically annotated datasets labeled for solidarity and its contestation, anti-solidarity. Our annotations use the textual material in social media posts to determine whether a post shows (anti-) solidarity with respect to relevant target groups. We achieve good agreement levels among experts and crowd annotators for a challenging novel NLP task. We further trained augmented BERT models whose performance is close to the agreement levels of the experts. We used these models for large-scale trend analysis of over 270 000 social media posts before and after the onset of the COVID-19 pandemic.

When examining migrant solidarity discourses directly related to COVID-19 (via corresponding keywords), we observe a sudden increase in COVID salience as the first measures are taken in Europe in mid- to end-March 2020, which then subsides rather quickly and remains permanently at low levels. In the migrant solidarity discourse which uses COVID keywords, levels of (anti-)solidarity depend on the subpopulation considered, but all of them show more solidarity in our data in the initial phase of the pandemic. A final observation of ours is that bursts in solidarity and anti-solidarity often go hand in hand, signaling issue salience and contestation. Focusing on only one of these phenomena can be deceptive as it suggests an unbalanced movement in society towards the political left or right.

We have shown the volatility of results to several factors, including the language users tweet in, whether they reveal their location or not, and the sampling strategy (with or without hashtags). We have observed that bursts of (anti-)solidarity may not be causally related to the pandemic, but may be related to other, co-occurring political events, necessitating a deeper look into the data to prevent misleading conclusions. Still, our findings provide robust evidence that migrant solidarity became increasingly salient and contested during the onset of the pandemic and that salience declined since late 2020, with tweet numbers falling just below pre-pandemic levels in summer 2021. During the same period, the share of anti-solidarity tweets increased in a sub-sample of COVID-19-related tweets, though solidarity tweets remained dominant. These findings highlight the importance of design choices, in particular long-term observation, pre- and post-crisis comparison and sampling in research interested in crisis related effects.

We hope this paper will serve to stimulate a wider discussion regarding research design and sampling strategy among scholars interested in (anti-)social online behavior during crises. As we have shown, using expressions of (anti-)solidarity towards migrants as examples of pro- and anti-social behavior during crises, design choices crucially inform findings and interpretation. In particular, we suggest to (1) establish a baseline against which dynamics of (anti-)solidarity can be compared; (2) consider dynamics of both pro- and anti-social behavior in light of issue salience vs. societal trends; (3) observe longer time spans, ideally capturing pre-, during-, and post-crisis phases; and (4) carefully reflect on sampling choices; preferably by comparing findings based on alternative sampling strategies. Whereas these suggestions are far from solving the methodological challenges that arise when trying to draw substantial conclusions based on online data, they can perhaps serve as a first step in creating awareness regarding the many challenges at hand.

References

[1] N. Fenton, Mediating solidarity, *Global Media and Communication*, vol. 4, no. 1, pp. 37–57, 2008.

[2] D. Margolin and W. Liao, The emotional antecedents of solidarity in social media crowds, *New Media & Society*, vol. 20, no. 10, pp. 3700–3719, 2018.

[3] S. Santhanam, V. Srinivasan, S. Glass, and S. Shaikh, I stand with you: Using emojis to study solidarity in crisis events, arXiv preprint arXiv: 1907.08326, 2019.

[4] Z. Tufekci, Social movements and governments in the digital age: Evaluating a complex landscape, *Journal of International Affairs*, vol. 68, no. 1, pp. 1–18, 2014.

[5] H. Silver, Social exclusion and social solidarity: Three paradigms, *International Labour Review*, vol. 133, nos. 5&6, pp. 531–578, 1994.

[6] D. Clarke, *Pro-Social and Anti-Social Behaviour*. New York, NY, USA: Routledge, Taylor & Francis Group, 2003.

[7] M. Mohler-Kuo, S. Dzemaili, S. Foster, L. Werlen, and S. Walitza, Stress and mental health among children/ adolescents, their parents, and young adults during the first COVID-19 lockdown in Switzerland, *International Journal of Environmental Research and Public Health*, vol. 18, no. 9, p. 4668, 2021.

[8] sCAN project, Hate speech trends during the COVID-19 pandemic in a digital and globalized age, https://scan-project.eu/resources-and-publications/#Covid-19, 2021.

[9] M. L. Williams, P. Burnap, A. Javed, H. Liu, and S. Ozalp, Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime, *The British Journal of Criminology*, vol. 60, no. 1, pp. 93–117, 2020.

[10] C. R. Seiter and N. S. Brophy, Social support and aggressive communication on social network sites during the COVID-19 pandemic, *Health communication*, doi: 10.1080/10410236.2021.1886399.

[11] J. B. Vieira, S. Pierzchajlo, S. Jangard, A. Marsh, and A. Olsson, Perceived threat and acute anxiety predict increased everyday altruism during the COVID-19 pandemic, http://doi.org/10.31234/osf.io/n3t5c, 2020.

[12] A. Stechemesser, L. Wenz, and A. Levermann, Corona crisis fuels racially profiled hate in social media networks, *EClinicalMedicine*, doi: 10.1016/j.eclinm.2020.100372.

[13] M. Comerford and L. Gerster, The rise of antisemitism online during the pandemic. A study of French and German content, Publications Office, https://op.europa.eu/en/publication-detail/-/publication/d73c833f-c34c-11eb-a925-01aa75ed71a1, 2021.

[14] O. Ozduzen and U. Korkut, Post-‘refugee crisis’ social

media: The unbearable lightness of sharing racist posts, *Discover Society*, https://archive.discoversociety.org/2020/09/02/post-refugee-crisis-social-media-the-unbearable-lightness-of-sharing-racist-posts/, 2020.

[15] J. Adam-Troian and S. C. Bagci, The pathogen paradox: Evidence that perceived COVID-19 threat is associated with both pro-and antiimmigrant attitudes, *International Review of Social Psychology*, vol. 34, no. 1, pp. 1–15, 2021.

[16] S. Masud, S. Dutta, S. Makkar, C. Jain, V. Goyal, A. Das, and T. Chakraborty, Hate is the new infodemic: A topic-aware modeling of hate speech diffusion on twitter, in *Proc. 2021 IEEE 37th International Conference on Data Engineering* (*ICDE*), Chania, Greece, 2021, pp. 504–515.

[17] M. R. Awal, R. Cao, S. Mitrovic, and R. K. -W. Lee, On analyzing antisocial behaviors amid COVID-19 pandemic, arXiv preprint arXiv: 2007.10712, 2020.

[18] A. Ils, D. Liu, D. Grunow, and S. Eger, Changes in European solidarity before and during COVID-19: Evidence from a large crowd- and expert-annotated Twitter dataset, in *Proc. 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (*Volume 1: Long Papers*), Online, 2021, pp. 1623–1637.

[19] L. Guadagno, Migrants and the COVID-19 pandemic: An initial analysis, https://publications.iom.int/system/files/pdf/mrs-60.pdf, 2020.

[20] S. Baglioni, O. Biosca, and T. Montgomery, Brexit, division, and individual solidarity: What future for Europe? Evidence from eight European countries, *American Behavioral Scientist*, vol. 63, no. 4, pp. 538–550, 2019.

[21] J. Gerhards, H. Lengfeld, Z. S. Ignácz, F. K. Kley, and M. Priem, *European Solidarity in Times of Crisis: Insights from a Thirteen-Country Survey*. London, UK: Routledge, 2019.

[22] S. Koos and V. Seibel, Solidarity with refugees across Europe. A comparative analysis of public support for helping forced migrants, *European Societies*, vol. 21, no. 5, pp. 704–728, 2019.

[23] C. Lahusen and M. T. Grasso, eds., *Solidarity in Europe: Citizens' Responses in Times of Crisis*. Cham, Switzerland: Springer International Publishing, 2018.

[24] M. Franceschelli, Global migration, local communities and the absent state: Resentment and resignation on the Italian island of Lampedusa, *Sociology* , vol. 54, no. 3, pp. 591–608, 2019.

[25] M. Gómez Garrido, M. A. Carbonero Gamundí, and A. Viladrich, The role of grassroots food banks in building political solidarity with vulnerable people, *European Societies*, vol. 21, no. 5, pp. 753–773, 2018.

[26] C. Heimann, S. Müller, H. Schammann, and J. Stürner, Challenging the nation-state from within: The emergence of transmunicipal solidarity in the course of the EU refugee controversy, *Social Inclusion*, vol. 7, no. 2, pp. 208–218, 2019.

[27] A. Nerghes and J. -S. Lee, Narratives of the refugee crisis: A comparative study of mainstream-media and twitter, *Media and Communication*, vol. 7, no. 2, pp. 275–288, 2019.

[28] S. Wallaschek, Solidarity in Europe in times of crisis, *Journal of European Integration*, vol. 41, no. 2, pp. 257–263, 2019.

[29] S. Wallaschek, Contested solidarity in the Euro crisis and Europe's migration crisis: A discourse network analysis, *Journal of European Public Policy*, vol. 27, no. 7, pp. 1034–1053, 2020.

[30] S. Wallaschek, The discursive construction of solidarity: Analysing public claims in Europe's migration crisis, *Political Studies*, vol. 68, no. 1, pp. 74–92, 2019.

[31] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[32] K. Binner and K. Scherschel, eds., *Fluchtmigration und Gesellschaft: Von Nutzenkalkulen, Solidaritat und Exklusion*. Arbeitsgesellschaft im Wandel, Beltz Juventa, 2019.

[33] G. Dragolov, Z. S. Ignácz, J. Lorenz, J. Delhey, K. Boehnke, and K. Unzicker, *Social Cohesion in the Western World: What Holds Societies Together: Insights from the Social Cohesion Radar*. Cham, Switzerland: Springer, 2016.

[34] S. R. Baker, A. Baksy, N. Bloom, S. J. Davis, and J. A. Rodden, Elections, political polarization, and economic uncertainty, *NBER Working Papers*, https://www.nber.org/system/files/working_papers/w27961/w27961.pdf, 2020.

[35] F. Nicoli, Hard-line Euroscepticism and the Eurocrisis: Evidence from a panel study of 108 elections across Europe, *JCMS*: *Journal of Common Market Studies*, vol. 55, no. 2, pp. 312–331, 2017.

[36] A. Bazo Vienrich and M. J. Creighton, What's left unsaid? In-group solidarity and ethnic and racial differences in opposition to immigration in the United States, *Journal of Ethnic and Migration Studies*, vol. 44, no. 13, pp. 2240–2255, 2017.

[37] D. Heerwegh, Mode differences between face-to-face and web surveys: An experimental investigation of data quality and social desirability effects, *International Journal of Public Opinion Research*, vol. 21, no. 1, pp. 111–121, 2009.

[38] A. L. Janus, The influence of social desirability pressures on expressed immigration attitudes, *Social Science Quarterly*, vol. 91, no. 4, pp. 928–946, 2010.

[39] M. Gangl and C. Giustozzi, The erosion of political trust in the great recession, *CORRODE Working Paper*, doi: 10.13140/RG.2.2.20930.07366.

[40] C. R. Sunstein, *#Republic: Divided Democracy in the Age of Social Media*. Princeton, NJ, USA: Princeton University Press, 2018.

[41] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, GoEmotions: A dataset of fine-grained emotions, in *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2020, pp. 4040–4054.

[42] K. Ding, J. Li, and Y. Zhang, Hashtags, emotions, and comments: A large-scale dataset to understand fine-grained social emotions to online topics, in *Proc. 2020 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Online, 2020, pp. 1376–1382.

[43] T. Haider, S. Eger, E. Kim, R. Klinger, and W.

Menninghaus, PO-EMO: Conceptualization, annotation, and modeling of aesthetic emotions in German and English poetry, in *Proc. 12th Language Resources and Evaluation Conference*, Marseille, France, 2020, pp. 1652–1663.

[44] C. Hutto and E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, pp. 216–225, 2014.

[45] L. A. M. Bostan and R. Klinger, An analysis of annotated corpora for emotion classification in text, in *Proc. 27th International Conference on Computational Linguistics*, Santa Fe, NM, USA, 2018, pp. 2104–2119.

[46] K. C. Fraser, I. Nejadgholi, and S. Kiritchenko, Understanding and countering stereotypes: A computational approach to the stereotype content model, in *Proc. 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (*Vol. 1: Long Papers*), Online, 2021, pp. 600–616.

[47] T. Beck, J. -U. Lee, C. Viehmann, M. Maurer, O. Quiring, and I. Gurevych, Investigating label suggestions for opinion mining in German COVID-19 social media, in *Proc. 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (*Vol. 1: Long Papers*), Online, 2021, pp. 1–13.

[48] T. Walter, C. Kirschner, S. Eger, G. Glavaš, A. Lauscher, and S. P. Ponzetto, Diachronic analysis of German parliamentary proceedings: Ideological shifts through the lens of political biases, in *Proc. 2021 ACM/IEEE Joint Conference on Digital Libraries*, Champaign, IL, USA, 2021, pp. 51–60.

[49] J. J. Jones, M. R. Amin, J. Kim, and S. Skiena, Stereotypical gender associations in language have decreased over time, *Sociological Science*, vol. 7, no. 1, pp. 1–35, 2020.

[50] RIAS, Antisemitic incidents in Germany 2020, annual report, https://report-antisemitism.de/en/documents/Antisemitic_incidents_in_Germany_Annual-Report_Federal_Association_RIAS_2020.pdf, 2021.

[51] F. Tahmasbi, L. Schild, C. Ling, J. Blackburn, G. Stringhini, Y. Zhang, and S. Zannettou, "Go Eat a Bat, Chang!": On the emergence of sinophobic behavior on web communities in the face of COVID-19, in *Proc. Web Conference 2021*, Ljubljana, Slovenia, 2021, pp. 1122–1133.

[52] S. A. Bin-Nashwan, M. Al-Daihani, H. Abdul-Jabbar, and L. H. A. Al-Ttaffi, Social solidarity amid the COVID-19 outbreak: Fundraising campaigns and donors' attitudes, *International Journal of Sociology and Social Policy*, vol. 42, nos. 3&4, pp. 232–247, 2020.

[53] S. Zajak, K. Stjepandić, and E. Steinhilper, Pro-migrant protest in times of COVID-19: Intersectional boundary spanning and hybrid protest practices, *European Societies*, vol. 23, no. sup1, pp. S172–S183, 2021.

[54] J. Devlin, M. -W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1* (*Long and Short Papers*), Minneapolis, MN, USA, 2019, pp. 4171–4186.

[55] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2020, pp. 8440–8451.

[56] J. He, J. Gu, J. Shen, and M. Ranzato, Revisiting self-training for neural sequence generation, presented at International Conference on Learning Representations （ICLR）2020，Addis Ababa, Ethiopia, 2020.

[57] S. Cao, N. Kitaev, and D. Klein, Multilingual alignment of contextual word representations, presented at 8th International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 2020.

[58] W. Zhao, G. Glavaš, M. Peyrard, Y. Gao, R. West, and S. Eger, On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation, in *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2020, pp. 1656–1671.

[59] W. Zhao, S. Eger, J. Bjerva, and I. Augenstein, Inducing language-agnostic multilingual representations, in *Proc. *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, Online, 2021, pp. 229–240.

[60] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier, in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 1135–1144.

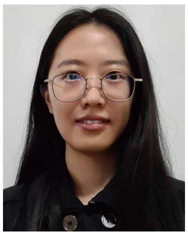
[61] S. F. Waterloo, S. E. Baumgartner, J. Peter, and P. M. Valkenburg, Norms of online expressions of emotion: Comparing Facebook, Twitter, Instagram, and Whatsapp, *New Media & Society* , vol. 20, no. 5, pp. 1813–1831, 2018.

[62] ECDC, European centre for disease prevention and control, data on country response measures to COVID-19, https://www.ecdc.europa.eu/sites/default/files/documents/response_graphs_data_2022-06-02.csv, 2020.

[63] I. Stewart, Y. Pinter, and J. Eisenstein, Si O no, que penses? Catalonian independence and linguistic identity on social media, in *Proc. 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, LA, USA, 2018, pp. 136–141.

Steffen Eger received the PhD degree in economics from Goethe University Frankfurt, Germany, in 2014. He was employed as a PostDoc at the Text Technology Lab in Frankfurt, Germany from 2014−2016 and the UKP Lab in Darmstadt, Germany during 2016−2018. During 2018−2022 he was the research group leader at TU Darmstadt. Currently, he is a stand-in professor at Bielefeld University, Germany. He has received the prestigious Heisenberg grant from the German Research Foundation (DFG) in 2022 and published widely in the Natural Language Processing (NLP) community. His current research interests include evaluation metrics for text generation systems and problems at the intersection of NLP, computational social sciences, and the digital humanities.

Daniela Grunow received the PhD degree in sociology (summa cum laude) from the University of Bamberg, Germany, in 2006. From 2006 to 2008, she was a postdoctoral associate at the Center for Research on Inequalities and the Life Course, Yale University, New Haven, CT, USA. During 2008−2012, she was a tenured faculty member at the Department of Sociology and Anthropology, University of Amsterdam, the Netherland. In 2011 she received the prestigious ERC starting grant from the European Research Council. She has been a full professor of sociology specializing in quantitative analyses of social change, at the Faculty of Social Sciences, Goethe University Frankfurt, Germany since 2013. She is the director of the Institute for Empirical-Analytical Research (InFER) at Goethe University and a spokesperson of the research group “ Reconfiguration and Internalization of Social Structure” (RISS, FOR5173), funded by the German Research Foundation. She is also a co-speaker of the Frankfurt division of the Research Institute Social Cohesion (RISC), funded by the German Federal Ministry of Education and Research (BMBF). She has published widely on social change with a focus on gender inequality and social cohesion.



Dan Liu is currently pursuing the master degree at Technical University Darmstadt, Germany. Her research interests include natural language processing and general computer science.