#MeTooMA: Multi-Aspect Annotations of Tweets Related to the MeToo Movement

Akash Gautam,[*1] **Puneet Mathur,**[*2] **Rakesh Gosangi,**[3] **Debanjan Mahata,**[3]
Ramit Sawhney,[4] **Rajiv Ratn Shah**[1]
[1]MIDAS, IIIT-Delhi {akash15011, rajivratn}@iiitd.ac.in,
[2]University of Maryland, College Park puneetm@cs.umd.edu,
[3]Bloomberg, New York, U.S.A. {rgosangi, dmahata}@bloomberg.net,
[4]Netaji Subhas Institute of Technology ramits.co@nsit.net.in

Abstract

In this paper, we present a dataset containing 9,973 tweets related to the MeToo movement that were manually annotated for five different linguistic aspects: relevance, stance, hate speech, sarcasm, and dialogue acts. We present a detailed account of the data collection and annotation processes. The annotations have a very high inter-annotator agreement (0.79 to 0.93 k-alpha) due to the domain expertise of the annotators and clear annotation instructions. We analyze the data in terms of geographical distribution, label correlations, and keywords. Lastly, we present some potential use cases of this dataset. We expect this dataset would be of great interest to psycholinguists, socio-linguists, and computational linguists to study the discursive space of digitally mobilized social movements on sensitive issues like sexual harassment.

Introduction

Over the last couple of years, the MeToo movement has facilitated several discussions about sexual abuse. Social media, especially Twitter, was one of the leading platforms where people shared their experiences of sexual harassment, expressed their opinions, and also offered support to victims. A large portion of these tweets was tagged with a dedicated hashtag #MeToo, and it was one of the leading trending topics in many countries. The movement was viral on social media, and the hashtag used over 19 million times[1] in a year.

The MeToo movement has been described as an essential development against the culture of sexual misconduct by many feminists, activists, and politicians. It is one of the primary examples of successful digital activism facilitated by social media platforms. The movement generated many conversations on stigmatized issues like sexual abuse and violence, which were not often discussed before because of the associated fear of shame or retaliation. This creates an opportunity for researchers to study how people express their opinion on a sensitive topic in an informal setting like social media. However, this is only possible if there are annotated datasets that explore different linguistic facets of such social media narratives.

Twitter served as a platform for many different types of narratives during the MeToo movement (Hosterman et al. 2018). It was used for sharing personal stories of abuse, offering support and resources to victims, and expressing support or opposition towards the movement (Lopez, Muldoon, and McKeown 2019). It was also used to allege individuals of sexual misconduct, refute such claims, and sometimes voice hateful or sarcastic comments about the campaign or individuals. In some cases, people also misused hashtag to share irrelevant or uninformative content. To capture all these complex narratives, we decided to curate a dataset of tweets related to the MeToo movement that is annotated for various linguistic aspects.

In this paper, we present a new dataset (MeTooMA[2]) that contains 9,973 tweets associated with the MeToo movement annotated for relevance, stance, hate speech, sarcasm, and dialogue acts. We introduce and annotate three new dialogue acts that are specific to the movement: Allegation, Refutation, and Justification. The dataset also contains geographical information about the tweets: from which country it was posted.

We expect this dataset would be of great interest and use to both computational and socio-linguists. For computational linguists, it provides an opportunity to model three new complex dialogue acts (allegation, refutation, and justification) and also to study how these acts interact with some of the other linguistic components like stance, hate, and sarcasm. For socio-linguists, it provides an opportunity to explore how a movement manifests in social media across multiple countries.

Related Datasets

Table 1 presents a summary of datasets that contain social media posts about sexual abuse and annotated for various labels.

- (Pandey et al. 2018) created a dataset of 2,500 tweets for the identification of malicious intent surrounding the

[*] equal contribution

[1] https://www.usatoday.com/story/news/2018/10/13/metoo-impact-hashtag-made-online/1633570002/

[2] The dataset can be found at https://doi.org/10.7910/DVN/JN4EYU.

Dataset	#Annotated Posts	Labels
(Pandey et al. 2018)	2500	*accusational, validation, sensational*
(Khatua, Cambria, and Khatua 2018)	1024	*assault at: workplace, educational institute, public place, home*
(Schrading et al. 2015)	18,336	*abuse, non-abuse*
(Chowdhury et al. 2019a)	5119	*recollection, non-recollection*
(Sharifirad and Jacovi 2019)	3240	*indirect sexism, casual sexism, physical sexism*
MeTooMA	9,937	*relevance, stance, hate speech, sarcasm, dialogue acts (allegation, justification, refutation)*

Table 1: Summary of related datasets.

cases of sexual assault. The tweets were annotated for labels like *accusational, validation, sensational*.

- (Khatua, Cambria, and Khatua 2018) collected 0.7 million tweets containing hashtags such as *#MeToo*, *#AlyssaMilano*, *#harassed*. The annotated a subset of 1024 tweets for the following assault-related labels: assault at the workplace by colleagues, assault at the educational institute by teachers or classmates, assault at public places by strangers, assault at home by a family member, multiple instances of assaults, or a generic tweet about sexual violence.
- (Schrading et al. 2015) created the Reddit Domestic Abuse Dataset, which contained 18,336 posts annotated for 2 classes, *abuse* and *non-abuse*.
- (Chowdhury et al. 2019a) presented a dataset consisting of 5119 tweets distributed into *recollection* and *non-recollection* classes. The tweet was annotated as *recollection* if it explicitly mentioned a personal instance of sexual harassment.
- (Sharifirad and Jacovi 2019) created a dataset with 3240 tweets labeled into three categories of sexism: *indirect sexism, casual sexism, physical sexism*.

SVAC (Sexual Violence in Armed Conflict) is another related dataset which contains reports annotated for six different aspects of sexual violence: *prevalence*, *perpetrators*, *victims*, *forms*, *location*, and *timing* (Sexual 2007).

Unlike all the datasets described above, which are annotated for a single group of labels, our dataset is annotated for **five different linguistic aspects**. It also has **more annotated samples** than most of its contemporaries.

Dataset

Data Collection

We focused our data collection over the period of October to December 2018 because October marked the one year anniversary of the MeToo movement. Our first step was to identify a list of countries where the movement was trending during the data collection period. To this end, we used Google's interactive tool named MeTooRisingWithGoogle[3], which visualizes search trends of the term "MeToo" across the globe. This helped us narrow down our query space to 16 countries.

We then scraped 500 random posts from online sexual harassment support forums to help identify keywords or phrases related to the movement [4]. The posts were first manually inspected by the annotators to determine if they were related to the MeToo movement. Namely, if they contained self-disclosures of sexual violence, relevant information about the events associated with the movement, references to news articles or advertisements calling for support for the movement. We then processed the relevant posts to extract a set of uni-grams and bi-grams with high tf-idf scores. The annotators further pruned this set by removing irrelevant terms resulting in a lexicon of 75 keywords. Some examples include: #Sexual Harassment, #TimesUp, #EveryDaySexism, assaulted, #WhenIwas, inappropriate, workplace harassment, groped, #NotOkay, believe survivors, #WhyIDidntReport.

We then used Twitter's public streaming API[5] to query for tweets from the selected countries, over the chosen three-month time frame, containing any of the keywords. This resulted in a preliminary corpus of 39,406 tweets. We further filtered this data down to include only English tweets based on tweet's *language* metadata field and also excluded short tweets (less than two tokens). Lastly, we de-duplicated the dataset based on the textual content. Namely, we removed all tweets that had more than 0.8 cosine similarity scores on the unaltered text in tf-idf space with any other tweet. We employed this de-duplication to promote more lexical diversity in the dataset. After this filtering, we ended up with a corpus of 9,973 tweets.

Table 2 presents the distribution of the tweets by country before and after the filtering process. A large portion of the samples is from India because the MeToo movement has peaked towards the end of 2018 in India. There are very few samples from Russia likely because of content moderation and regulations on social media usage in the country[6]. Figure 1 gives a geographical distribution of the curated dataset.

Due to the sensitive nature of this data, we have decided to remove any personal identifiers (such as names, locations, and hyperlinks) from the examples presented in this paper. We also want to caution the readers that some of the examples in the rest of the paper, though censored for

[3] https://metoorising.withgoogle.com/

[4] We scraped data from the discussion forums on the websites of two non-profit organizations (pandys and isurvive), which provide support and resources to survivors of abuse.

[5] https://www.tweepy.org/

[6] https://time.com/5636107/metoo-russia-womens-rights/

Country	#Tweets	#Filtered Tweets
India	20,112	5,082
USA	8,943	2,773
United Kingdom	4,350	1,334
France	1,120	347
Australia	542	153
South Africa	1,085	103
Japan	830	13
Kenya	696	15
UAE	540	51
New Zealand	248	38
Iran	325	7
Canada	324	24
Sweden	139	20
Spain	62	9
Austria	88	2
Russia	42	2
Total	**39,406**	**9,973**

Table 2: Distribution of tweets by the country.

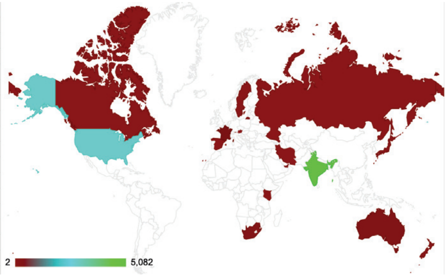


Figure 1: Choropleth world map recording tweet frequency.

profanity, contain offensive language and express a harsh sentiment.

Annotation Task

We chose against crowd-sourcing the annotation process because of the sensitive nature of the data and also to ensure a high quality of annotations. We employed three domain experts who had advanced degrees in clinical psychology and gender studies. The annotators were first provided with the guidelines document, which included instructions about each task, definitions of class labels, and examples. They studied this document and worked on a few examples to familiarize themselves with the annotation task. They also provided feedback on the document, which helped us refine the instructions and class definitions. The annotation process was broken down into five sub-tasks: for a given tweet, the annotators were instructed to identify relevance, stance, hate speech, sarcasm, and dialogue act. An important consideration was that the sub-tasks were not mutually exclusive, implying that the presence of one label did not consequently mean an absence of any.

Task 1: Relevance Here the annotators had to determine if the given tweet was relevant to the MeToo movement. Relevant tweets typically include personal opinions (either positive or negative), experiences of abuse, support for victims, or links to MeToo related news articles. Following are examples of a *relevant* tweet:

> *Officer [name] could be kicked out of the force after admitting he groped a woman at [place] festival last year. His lawyer argued saying the constable shouldn't be punished because of the #MeToo movement. #notokay #sexualabuse.*

and an *irrelevant* tweet:

> *Had a bit of break. Went to the beautiful Port [place] and nearby areas. Absolutely stunning as usual. #beautiful #MeToo #Australia #auspol [URL].*

We expect this relevance annotation could serve as a useful filter for downstream modeling.

Task 2: Stance Stance detection is the task of determining if the author of a text is in favor or opposition of a particular target of interest (Augenstein et al. 2016; Mohammad et al. 2016). Stance helps understand public opinion about a topic and also has downstream applications in information extraction, text summarization, and textual entailment (Sobhani 2017). We categorized stance into three classes: Support, Opposition, Neither. Support typically included tweets that expressed appreciation of the MeToo movement, shared resources for victims of sexual abuse, or offered empathy towards victims. Following is an example of a tweet with a *Support* stance:

> *Opinion: #MeToo gives a voice to victims while bringing attention to a nationwide stigma surrounding sexual misconduct at a local level.[URL]. This should go on.*

On the other hand, Opposition included tweets expressing dissent over the movement or demonstrating indifference towards the victims of sexual abuse or sexual violence. An example of a *Opposition* tweet is shown below:

> *The double standards and selective outrage make it clear that feminist concerns about power imbalances in the workplace aren't principles but are tools to use against powerful men they hate and wish to destroy. #fakefeminism. #men.*

Task 3: Hate Speech Detection of hate speech in social media has been gaining interest from NLP researchers lately (Waseem and Hovy 2016; Badjatiya et al. 2017). Our annotation scheme for hate speech is based on the work of (Basile et al. 2019). For a given tweet, the annotators first had to determine if it contained any hate speech. If the tweet was hateful, they had to identify if the hate was *Directed* or *Generalized*. Directed hate is targeted at a particular individual or entity, whereas Generalized hate is targeted at larger groups that belonged to a particular ethnicity, gender, or sexual orientation. Following are examples of tweets with *Directed* hate:

> *[username] were lit minus getting f*c*i*g mouth raped by some drunk chick #MeToo (nobody cares because I'm a male) [URL]*

and *Generalized* hate:

*For the men who r asking "y not then, y now?", u guys will still doubt her & harass her even more for y she shared her story immediately no matter what! When your sister will tell her childhood story to u one day, I challenge u guys to ask "y not then, y now?" #Metoo [username] [URL] #a**holes.*

Task 4: Sarcasm Sarcasm detection has also become a topic of interest for computational linguistics over the last few years (Bamman and Smith 2015; Rajadesingan, Zafarani, and Liu 2015) with applications in areas like sentiment analysis and affective computing. Sarcasm was an integral part of the MeToo movement. For example, many women used the hashtag #NoWomanEver to sarcastically describe some of their experiences with harassment[7]. We instructed the annotators to identify the presence of any sarcasm in a tweet, either about the movement or about an individual or entity. Following is an example of a sarcastic tweet:

# was pound before it was a hashtag. If you replace hashtag with the pound in the #metoo, you get, pound me too. Does that apply to [name].

Task 5: Dialogue Acts A dialogue act is defined as the function of a speaker's utterance during a conversation (McTear, Callejas, and Griol 2016), for example, question, answer, request, suggestion, etc. Dialogue Acts have been extensive studied in spoken (Ang, Liu, and Shriberg 2005) and written (Kim, Cavedon, and Baldwin 2010) conversations and have lately been gaining interest in social media (Zarisheva and Scheffler 2015). In this task, we introduced three new dialogue acts that are specific to the MeToo movement: Allegation, Refutation, and Justification.

Allegation: This category includes tweets that allege an individual or a group of sexual misconduct. The tweet could either be personal opinion or text summarizing allegations made against someone (Hutchings 2012). The annotators were instructed to identify if the tweet includes the hypothesis of allegation based on a first-hand account or a verifiable source confirming the allegation. Following is an example of a tweet that qualifies as an Allegation:

More women accuse [name] of grave sexual misconduct...twitter seethes with anger. #MeToo #pervert.

Refutation: This category contains tweets where an individual or an organization is denying allegations with or without evidence. Following is an example of a Refutation tweet:

She is trying to use the #MeToo movement to settle old scores says [name1] after [name2] levels sexual assault allegations against him.

Justification: The class includes tweets where the author is justifying their actions. These could be alleged actions in the real world (e.g., an allegation of sexual misconduct) or some action performed on twitter (e.g., supporting someone who was alleged of misconduct). Following is an example of a tweet that would be tagged as Justification:

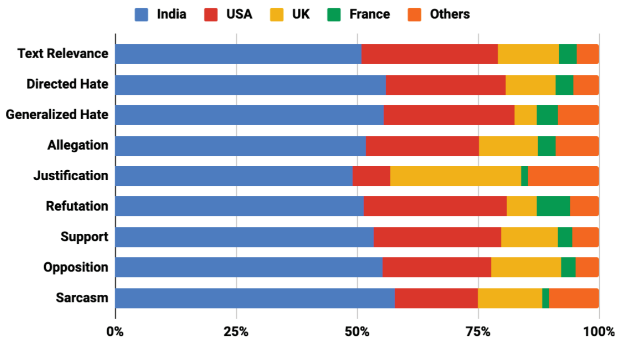


Figure 2: Geographical distribution of various class labels.

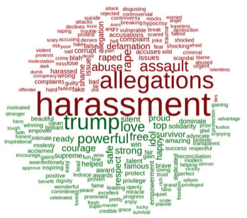


Figure 3: Word cloud representation of the dataset: font size is proportional to the frequency of a term. The words are organized and color-coded based on the NRC sentiment lexicon: positive sentiment (green + bottom half), negative sentiment (red + top half).

I actually did try to report it, but he and of his friends got together and lied to the police about it. #WhyIDidNotReport.

Dataset Analysis

This section includes descriptive and quantitative analysis performed on the dataset.

Inter-annotator agreement

We evaluated inter-annotator agreements using Krippendorff's alpha (K-alpha) (Krippendorff 2011). K-alpha, unlike simple agreement measures, accounts for chance correction and class distributions and can be generalized to multiple annotators. Table 4 summarizes the K-alpha measures for all the annotation tasks. We observe very strong agreements for most of the tasks with a maximum of 0.92 for the relevance task. The least agreement observed was for the hate speech task at 0.78. Per recommendations in (Artstein and Poesio 2008), we conclude that these annotations are of good quality. We chose a straightforward approach of majority decision for label adjudication: if two or more annotators agreed on assigning a particular class label. In cases of discrepancy, the labels were adjudicated manually by the authors. Table 5 shows a distribution of class labels after adjudication.

[7] https://www.good.is/articles/maura-quint-twitter-sexual-assault

Directed Hate	**SAGE**	**Generalized Hate**	**SAGE**
f*ck	3.36	hate	3.21
f*cking	3.04	lie	2.95
hijab	2.84	predators	2.92
bullshit	2.77	nuns	2.91
blog	2.70	grop	2.91
Allegation	**SAGE**	**Justification**	**SAGE**
accuse	1.45	organisation	0.57
bob	1.45	told	0.56
flopping	1.40	discuss	0.56
aces	1.40	violent	0.55
corrupt	1.35	shocked	0.51
Support	**SAGE**	**Opposition**	**SAGE**
fund	0.80	mocks	2.47
reconciliation	0.66	tweet	2.19
diversity	0.62	practice	2.19
protect	0.62	feminism	2.11
welcome	0.59	minister	2.11
Refutation	**SAGE**	**Sarcasm**	**SAGE**
baseless	3.63	lol	2.74
wild	3.59	gonna	2.71
center	3.46	trouble	2.71
denies	3.17	ooh	2.41
threatens	3.07	xoxo	2.20

Table 3: Top five phrases learned by SAGE Topic model for the all the labels

Task	Krippendorff's α
Relevance	0.92
Stance	0.90
Hate speech	0.78
Sarcasm	0.80
Allegation	0.86
Refutation	0.83
Justification	0.79

Table 4: Inter-annotator agreements for all the annotation tasks.

Geographical Distribution

Figure 2 presents a distribution of all the tweets by their country of origin. As expected, a large portion of the tweets across all classes are from India, which is consistent with Table 2. Interestingly, the US contributes a comparatively smaller proportion of tweets to the Justification category, and likewise, UK contributes a lower portion of tweets to the Generalized Hate category. Further analysis is necessary to establish if these observations are statistically significant.

Label Correlations

We conducted a simple experiment to understand the linguistic similarities (or lack thereof) for different pairs of class labels both within and across tasks. To this end, for each pair of labels, we converted the data into its tf-idf representation and then estimated Pearson, Spearman, and Kendall Tau correlation coefficients and also the corresponding p values. The results are summarized in Table 6. Overall, the correlation values seem to be on a lower end with maximum Pearson's correlation value obtained for the label pair *Justification - Support*, maximum Kendall Tau's correlation for *Allegation - Support*, and maximum Spearman's correlation for *Directed Hate - Generalized Hate*. The correlations are statistically significant ($p < 0.05$) for three pairs of class labels: *Directed Hate - Generalized Hate, Directed Hate - Opposition, Sarcasm - Opposition*. Sarcasm and Allegation also have statistically significant p values for Pearson and Spearman correlations.

Task	Label	#Samples	%
Relevance	Relevant	7,249	72.8%
Stance	Support	3,074	30.9%
	Opposition	743	7.4%
Hate Speech	Directed	419	4.21%
	Generalized	281	2.8%
Sarcasm	Sarcastic	220	2.2%
Dialogue Acts	Allegation	578	5.78%
	Justification	292	2.9%
	Refutation	216	2.1%

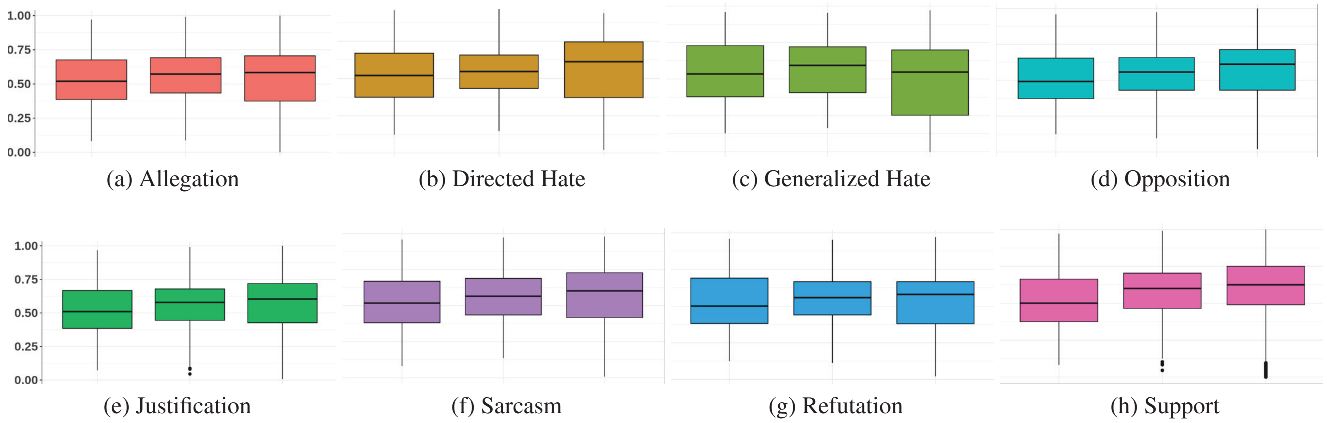
Table 5: Distribution of class labels for all tasks.

Label pair	PCC	p-PCC	KCC	p-KCC	SCC	p-SCC
Directed Hate - Generalized Hate	0.049	**0.0432**	0.268	**0.0021**	**0.477**	**0.0344**
Directed Hate - Sarcasm	0.052	0.0731	0.252	0.0521	0.258	0.0623
Directed Hate - Allegation	0.045	0.0832	0.244	0.0712	0.252	0.0523
Directed Hate - Justification	0.049	0.0661	0.413	0.0053	0.381	0.0503
Directed Hate - Refutation	0.054	0.5391	0.314	0.0044	0.322	0.0712
Directed Hate - Support	0.073	0.0882	0.042	0.0621	0.303	0.0032
Directed Hate - Opposition	0.061	**0.0022**	0.314	**0.0450**	0.322	**0.0433**
Generalized Hate - Sarcasm	0.062	0.0233	0.260	0.0051	0.265	0.0421
Generalized Hate - Allegation	0.059	0.0644	0.266	0.0260	0.271	0.0345
Generalized Hate - Justification	0.034	0.0633	0.271	0.0532	0.281	0.0611
Generalized Hate -Refutation	0.051	0.0821	0.223	0.0558	0.230	0.0031
Generalized Hate - Support	0.028	0.6820	0.325	0.0621	0.355	0.0652
Generalized Hate - Opposition	0.068	0.0239	0.320	0.0030	0.341	0.0532
Sarcasm - Allegation	0.045	0.0471	0.244	0.0613	0.202	0.0072
Sarcasm - Justification	0.061	0.0891	0.281	0.0401	0.013	0.0014
Sarcasm - Refutation	0.035	0.0772	0.243	0.0023	0.221	0.0833
Sarcasm - Support	0.064	0.0514	0.233	0.0080	0.259	0.0041
Sarcasm - Opposition	0.062	**0.0034**	0.271	**0.0430**	0.362	**0.0332**
Allegation - Justification	0.053	0.0499	0.251	0.0031	0.262	0.0023
Allegation - Refutation	0.062	0.0344	0.280	0.0421	0.281	0.0014
Allegation - Support	0.027	0.6711	**0.467**	0.0631	0.003	0.0779
Allegation - Opposition	0.574	0.6533	0.359	0.0231	0.205	0.0702
Justification - Refutation	0.443	0.6688	0.226	0.0711	0.226	0.0244
Justification - Support	**0.742**	0.7121	0.311	0.0093	0.311	0.0261
Justification - Opposition	0.734	0.0429	0.326	0.0201	0.385	0.0342
Refutation - Support	0.562	0.0822	0.237	0.0718	0.252	0.0522
Refutation - Opposition	0.651	0.0633	0.433	0.0433	0.043	0.0521
Support - Opposition	0.234	0.0533	0.249	0.7213	0.272	0.0852

Table 6: Correlation coefficients and p-values for each pair of labels in the dataset.

Keywords

We used SAGE (Eisenstein, Ahmed, and P Xing 2011), a topic modeling method, to identify keywords associated with the various class labels in our dataset. SAGE is an unsupervised generative model that can identify words that distinguish one part of the corpus from rest. For our keyword analysis, we removed all the hashtags and only considered tokens that appeared at least five times in the corpus, thus ensuring they were representative of the topic. Table 3 presents the top five keywords associated with each class and also



(a) Allegation (b) Directed Hate (c) Generalized Hate (d) Opposition

(e) Justification (f) Sarcasm (g) Refutation (h) Support

Figure 4: Arousal, Dominance, and Valence scores for all class labels based on NRC VAD lexicon for each of the labels. The first box presents arousal score, the second one dominance score, and the third one valence dimension.

their salience scores. Though *Directed* and *Generalized* hate are closely related topics, there is not much overlap between the top 5 salient keywords suggesting that there are linguistic cues to distinguish between them. The word `predators` is strongly indicative of *Generalized Hate*, which is intuitive because it is a term often used to describe people who were accused of sexual misconduct. The word `lol` being associated with *Sarcasm* is also reasonably intuitive because of sarcasm's close relation with humor.

Sentiment Analysis

Figure 3 presents a word cloud representation of the data where the colors are assigned based on the NRC emotion lexicon (Mohammad and Turney 2013), green for positive and red for negative. We also analyzed all the classes in terms of Valence, Arousal, and Dominance using the NRC VAD lexicon (Mohammad 2018). The results are summarized in Figure 4. Of all the classes, *Directed-Hate* has the largest valence spread, which is likely because of the extreme nature of the opinions expressed in such tweets. The spread for the dominance is fairly narrow for all class labels, with the median score slightly above 0.5, suggesting a slightly dominant nature exhibited by the authors of the tweets.

Discussion

This paper introduces a new dataset containing tweets related to the #MeToo movement. It may involve opinions over socially stigmatized issues or self-reports of distressing incidents. Therefore, it is necessary to examine the social impact of this exercise, the ethics of the individuals concerned with the dataset, and it's limitations.

Mental health implications: This dataset open source posts curated by individuals who may have undergone instances of sexual exploitation in the past. While we respect and applaud their decision to raise their voices against their exploitation, we also understand that their revelations may have been met with public backlash and apathy in both the virtual as well as the real world. In such situations, where the social reputation of both accuser and accused may be under threat, mental health concerns become very important[8]. As survivors recount their horrific episodes of sexual harassment, it becomes imperative to provide them with therapeutic care (Fredriksen-Goldsen et al. 2014; Chowdhury et al. 2019b) as a safeguard against mental health hazards. Such measures, if combined with the integration of mental health assessment tools in social media platforms, can make victims of sexual abuse feel more empowered and self-contemplative towards their revelations.

Use of MeTooMA dataset for population studies: We would like to mention that there have been no attempts to conduct population-centric analysis on the proposed dataset. The analysis presented in this dataset should be seen as a proof of concept to examine the instances of the #MeToo movement on Twitter. The authors acknowledge that learning from this dataset cannot be used as-is for any direct social interventions. Network sampling of real-world users for any experimental work beyond this dataset would require careful evaluation beyond the observational analysis presented herein. Moreover, the findings could be used to assist already existing human knowledge. Experiences of the affected communities should be recorded and analyzed carefully, which could otherwise lead to social stigmatization, discrimination, and societal bias. Enough care has been ensured so that this work does not come across as trying to target any specific individual for their personal stance on the issues pertaining to the social theme at hand. The authors do not aim to vilify individuals accused in the #MeToo cases in any manner. Our work tries to bring out general trends that may help researchers develop better techniques to understand mass unorganized virtual movements.

[8] https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(18)30991-7/fulltext

Effect on marginalized communities: The authors recognize the impact of the #MeToo movement on socially stigmatized populations like LGBTQIA+. The #MeToo movement provided such individuals with the liberty to express their notions about instances of sexual violence and harassment[9]. The movement acted as a catalyst towards implementing social policy changes to benefit the members of these communities[10]. Hence, it is essential to keep in mind that any experimental work undertaken on this dataset should try to minimize the biases against the minority groups which might get amplified in cases of a sudden outburst of public reactions over sensitive media discussions.

Limitations of individual consent: Considering the mental health aspects of the individuals concerned, social media practitioners should vary in making automated interventions to aid the victims of sexual abuse as some individuals might not prefer to disclose their sexual identities or notions. Concerned social media users might also repeal their social media information if found out that their personal information may be potentially utilized for computational analysis. Hence, it is imperative to seek subtle individual consent before trying to profile authors involved in online discussions to uphold personal privacy.

Use Cases

The authors would like to formally propose some ideas on possible extensions of the proposed dataset:

- The rise of online **hate speech** and its related behaviors like cyber-bullying has been a hot topic of research in gender studies (Djuric et al. 2015). Our dataset could be utilized for extracting actionable insights and virtual dynamics to identify gender roles for analyzing sexual abuse revelations similar to (Yuce et al. 2014).
- The dataset could be utilized by psycholinguistics for extracting contextualized lexicons to examine how influential people are portrayed on public platforms in events of mass social media movements (Field, Bhat, and Tsvetkov 2019). Interestingly, such analysis may help linguists determine the **power dynamics of authoritative people** in terms of perspective and sentiment through campaign modeling.
- Marginalized voices affected by mass social movements can be studied through **polarization analysis** on graph-based simulations of the social media networks. Based on the data gathered from these nodes, community interactions could be leveraged to identify indigenous issues pertaining to societal unrest across various sections of the society(Rho, Mark, and Mazmanian 2018).
- **Challenge Proposal**: The authors of the paper would like to extend the present work as a challenge proposal for building computational semantic analysis systems aimed at online social movements. In contrast to already available datasets and existing challenges, we propose tasks on detecting hate speech, sarcasm, stance, and relevancy that will be more focused on social media activities surrounding revelations of sexual abuse and harassment. The tasks may utilize the message-level text, linked images, tweet-level metadata and user-level interactions to model systems that are **F**air, **A**ccountable, **I**nterpretable and **R**esponsible (FAIR).

Research ideas emerging from this work should not be limited to the above discussion. If needed, supplementary data required to enrich this dataset can be collected utilizing Twitter API and *JSON* records for exploratory tasks beyond the scope of the paper.

Conclusion

In this paper, we presented a new dataset annotated for five different linguistic aspects: relevance, stance, hate speech, sarcasm, and dialogue acts. To our knowledge, there are no datasets out there that provide annotations across so many different dimensions. This allows researchers to perform various multi-label and multi-aspect classification experiments. Additionally, researchers could also address some interesting questions on how different linguistic components influence each other: e.g., does understanding one's stance help in better prediction of hate speech?

In addition to these exciting computational challenges, we expect this data could be useful for socio and psycholinguists in understanding the language used by victims when disclosing their experiences of abuse. Likewise, they could analyze the language used by alleged individuals in justifying their actions. It also provides a chance to examine the language used to express hate in the context of sexual abuse.

In the future, we would like to propose challenge tasks around this data where the participants will have to build computational models to capture all the different linguistic aspects that were annotated. We expect such a task would drive researchers to ask more interesting questions, find limitations of the dataset, propose improvements, and provide interesting insights.

Acknowledgments

Rajiv Ratn Shah is partly supported by the Infosys Center for AI, IIIT Delhi.

References

Ang, J.; Liu, Y.; and Shriberg, E. 2005. Automatic dialog act segmentation and classification in multiparty meetings. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, I–1061. IEEE.

Artstein, R., and Poesio, M. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34(4):555–596.

Augenstein, I.; Rocktäschel, T.; Vlachos, A.; and Bontcheva, K. 2016. Stance detection with bidirectional conditional encoding. *arXiv preprint arXiv:1606.05464*.

Badjatiya, P.; Gupta, S.; Gupta, M.; and Varma, V. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of*

[9] https://rewire.news/article/2018/10/09/for-lgbtq-youth-metoo-is-not-a-heteronormative-issue/

[10] https://www.reuters.com/article/us-lgbt-rights-twitter/mequeer-takes-twitter-by-storm-as-lgbt-community-cries-metoo-idUSKCN1L71WW

the 26th International Conference on World Wide Web Companion, 759–760. International World Wide Web Conferences Steering Committee.

Bamman, D., and Smith, N. A. 2015. Contextualized sarcasm detection on twitter. In *Ninth International AAAI Conference on Web and Social Media*.

Basile, V.; Bosco, C.; Fersini, E.; Nozza, D.; Patti, V.; Rangel Pardo, F. M.; Rosso, P.; and Sanguinetti, M. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 54–63. Minneapolis, Minnesota, USA: Association for Computational Linguistics.

Chowdhury, A. G.; Sawhney, R.; Mathur, P.; Mahata, D.; and Shah, R. R. 2019a. Speak up, fight back! detection of social media disclosures of sexual harassment. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, 136–146.

Chowdhury, A. G.; Sawhney, R.; Shah, R.; and Mahata, D. 2019b. # youtoo? detection of personal recollections of sexual harassment on social media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2527–2537.

Djuric, N.; Zhou, J.; Morris, R.; Grbovic, M.; Radosavljevic, V.; and Bhamidipati, N. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, 29–30. ACM.

Eisenstein, J.; Ahmed, A.; and P Xing, E. 2011. Sparse additive generative models of text.

Field, A.; Bhat, G.; and Tsvetkov, Y. 2019. Contextual affective analysis: A case study of people portrayals in online# metoo stories. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 158–169.

Fredriksen-Goldsen, K. I.; Hoy-Ellis, C. P.; Goldsen, J.; Emlet, C. A.; and Hooyman, N. R. 2014. Creating a vision for the future: Key competencies and strategies for culturally competent practice with lesbian, gay, bisexual, and transgender (lgbt) older adults in the health and human services. *Journal of gerontological social work* 57(2-4):80–107.

Hosterman, A. R.; Johnson, N. R.; Stouffer, R.; and Herring, S. 2018. Twitter, social support messages, and the# metoo movement. *The Journal of Social Media in Society* 7(2):69–91.

Hutchings, C. 2012. Commercial use of facebook and twitter–risks and rewards. *Computer Fraud & Security* 2012(6):19–20.

Khatua, A.; Cambria, E.; and Khatua, A. 2018. Sounds of silence breakers: Exploring sexual violence on twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 397–400. IEEE.

Kim, S. N.; Cavedon, L.; and Baldwin, T. 2010. Classifying dialogue acts in one-on-one live chats. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 862–871. Association for Computational Linguistics.

Krippendorff, K. 2011. Computing krippendorff's alpha-reliability.

Lopez, K. J.; Muldoon, M. L.; and McKeown, J. K. 2019. One day of# feminism: Twitter as a complex digital arena for wielding, shielding, and trolling talk on feminism. *Leisure Sciences* 41(3):203–220.

McTear, M. F.; Callejas, Z.; and Griol, D. 2016. *The conversational interface*, volume 6. Springer.

Mohammad, S. M., and Turney, P. D. 2013. Crowdsourcing a word-emotion association lexicon. 29(3):436–465.

Mohammad, S.; Kiritchenko, S.; Sobhani, P.; Zhu, X.; and Cherry, C. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 31–41.

Mohammad, S. M. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*.

Pandey, R.; Purohit, H.; Stabile, B.; and Grant, A. 2018. Distributional semantics approach to detect intent in twitter conversations on sexual assaults. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 270–277. IEEE.

Rajadesingan, A.; Zafarani, R.; and Liu, H. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 97–106. ACM.

Rho, E. H. R.; Mark, G.; and Mazmanian, M. 2018. Fostering civil discourse online: Linguistic behavior in comments of# metoo articles across political perspectives. *Proceedings of the ACM on Human-Computer Interaction* 2(CSCW):147.

Schrading, N.; Alm, C. O.; Ptucha, R.; and Homan, C. 2015. An analysis of domestic abuse discourse on reddit. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2577–2583.

Sexual, V. 2007. Sexual violence in armed conflict.

Sharifirad, S., and Jacovi, A. 2019. Learning and understanding different categories of sexism using convolutional neural network's filters. In *Proceedings of the 2019 Workshop on Widening NLP*, 21–23.

Sobhani, P. 2017. *Stance detection and analysis in social media*. Ph.D. Dissertation, Université d'Ottawa/University of Ottawa.

Waseem, Z., and Hovy, D. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, 88–93.

Yuce, S. T.; Agarwal, N.; Wigand, R. T.; Lim, M.; and Robinson, R. S. 2014. Bridging women rights networks: Analyzing interconnected online collective actions. *Journal of Global Information Management (JGIM)* 22(4):1–20.

Zarisheva, E., and Scheffler, T. 2015. Dialog act annotation for twitter conversations. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 114–123.