

# Multitask Learning for Emotionally Analyzing Sexual Abuse Disclosures

Ramit Sawhney<sup>†</sup>, Puneet Mathur<sup>‡</sup>, Taru Jain<sup>†</sup>, Akash Kumar Gautam<sup>†</sup>, Rajiv Ratn Shah<sup>†</sup>

<sup>†</sup> Department of Computer Engineering, IIT-Delhi

{ramits, akash15011, rajivrtn}@iiitd.ac.in

<sup>‡</sup> University of Maryland, College Park

puneetm@cs.umd.edu

## Abstract

The #MeToo movement on social media platforms initiated discussions over several facets of sexual harassment in our society. Prior work by the NLP community for automated identification of the narratives related to sexual abuse disclosures barely explored this social phenomenon as an independent task. However, emotional attributes associated with textual conversations related to the #MeToo social movement are complexly intertwined with such narratives. We formulate the task of identifying narratives related to the sexual abuse disclosures in online posts as a joint modeling task that leverages their emotional attributes through multitask learning. Our results demonstrate that positive knowledge transfer via context-specific shared representations of a flexible cross-stitched parameter sharing model helps establish the inherent benefit of jointly modeling tasks related to sexual abuse disclosures with emotion classification from the text in homogeneous and heterogeneous settings. We show how for more domain-specific tasks related to sexual abuse disclosures such as sarcasm identification and dialogue act (refutation, justification, allegation) classification, homogeneous multitask learning is helpful, whereas for more general tasks such as stance and hate speech detection, heterogeneous multitask learning with emotion classification works better.<sup>1</sup>

## 1 Introduction

The #MeToo movement<sup>2</sup> was started as an initiative to empower women against long-standing issues related to sexual abuse at workplaces, public spaces, and private organizations (McKenna and Chughtai, 2020). The usage of a dedicated hashtag #MeToo on media platforms signified a social support system for women from different sections

<sup>1</sup>Code & Implementation: <https://github.com/midas-research/metoo-mtl-naacl>

<sup>2</sup><https://metoomvmt.org/>

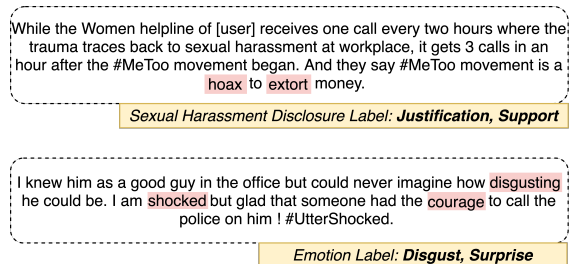


Figure 1: Examples showing the relationship between tweets annotated for sexual harassment disclosure (top) and emotion recognition (bottom). Colors highlight token level attention assigned by BERTweet.

of society. The movement initiated discussions on many socially stigmatized issues that were missing from the virtual space (Clark-Parsons, 2019). Such conversations invited various reactions on the web, involving *support* to the cause of the movement and even outright bullying. While many users took part in the vilification of the survivors, the movement also saw *opposition* by factions of the society that felt threatened by the impact of social media in raising awareness about the scale of everyday sexual harassment faced by women in workplaces and institutions (Tambe, 2018). In many instances, the public disclosures of survivor-narrated incidents involved widespread use of *hate-language* and online trolling, both against the victims and alleged oppressors (Franks, 2019). The #MeToo movement also led to people coming out with *allegations, refutations, and justifications* about traumatic experiences as they transitioned to active participants in the mainstream conversation (Gautam et al., 2020). A closer look at the online posts about the #MeToo movement revealed that *sarcasm* was often used as a thin veil in such discussions to humorously mask disapproval, wit, and personal attacks (Sandhu et al., 2019).

The complex narratives present in the conversations on stigmatized issues like sexual abuse create an opportunity for researchers to study how people

express their opinions on a sensitive topic in an informal social setting. It also offers a chance to social media regulators for fostering social inclusion, community integration, and improving the individual perception of being supported by others. This paper aims at categorizing the posts related to the #MeToo movement on the basis of stance (support or opposition), hate-speech, sarcasm, and dialogue acts (allegation, refutation, or justification of sexual misconduct). We focus our analysis on a publicly available dataset that is created in the backdrop of mass instances of sexual harassment disclosures and includes nuanced labels to identify accompanying linguistic behaviors.

Existing literature has emphasized that the text’s emotional attributes have a high correlation with dialogue narratives describing instances of sexual harassment (Lane and Hedin, 2020). Prior works (Anzovino et al., 2018; Sharifirad et al., 2018) have mostly focused on label specific detection of linguistic narratives related to sexual harassment disclosures in isolation by exploiting lexical features (Chowdhury et al., 2019; Karlekar and Bansal, 2018). However, subtle intricacies present in the discussion of sexual abuse disclosures often reflect the speaker’s affective and psychological state, which are overlooked by feature-engineered models. For instance, part (a) of Figure 1 shows a tweet expressing *support* towards the #MeToo movement but in a tone that might be difficult for naive neural learning models to capture without context. Part (b) of Figure 1 presents a tweet in which the author has an initial positive outlook, which later reverses to disgust for the subject. The lack of context about the event and contrasting qualifications describing the oppressor makes the correct classification of the sexual harassment disclosure label extremely challenging for traditional classifiers without emotional labels’ additional supervision.

Moreover, apart from their inherent complexity, conversations related to the #MeToo movement also pose a challenge of emotional ambiguity. This work is the first attempt at joint modeling of narratives related to sexual abuse disclosures and emotion classification to learn the patterns of their interaction via parameter sharing techniques offered by Multitask Learning (MTL). The affective features, which result from a joint learning setup through shared parameters, will encompass the text’s emotional content that is likely to be predictive of narratives corresponding to sexual abuse disclo-

ures. More specifically, we formulate an MTL framework for multi-label classification of narratives related to sexual abuse disclosures (stance, hate-speech, sarcasm, dialogue acts) and emotional classification in the context of the #MeToo movement. MTL (Caruana, 1997) allows two or more related tasks to be learned jointly. This facilitates the transfer of inductive bias and better generalization across related tasks on account of shared representations of linguistic features.

**Contributions** We experiment with MTL architectures employing a flexible cross-stitched parameter sharing method that benefits from both hard-parameter sharing and soft parameter sharing through a gated mechanism using a weighted summation (Section 4). Hard parameter sharing allows for sharing lower-level word representations, and soft parameter sharing permits the sharing of task-specific networks. We explore two flavors of multitask learning: (i) *Homogeneous MTL* - Intra-domain MTL between related tasks of sexual abuse disclosure narratives, and (ii) *Heterogeneous MTL* - cross-domain MTL between pairs of tasks in emotion classification and narratives of sexual abuse disclosure (Section 5.2). Our results demonstrate that both Homogeneous and Heterogeneous MTL setups outperform the Single Task Learning (STL) technique across various tasks (Section 6). Further, we conduct a qualitative analysis of several samples to analyze the benefit of joint training of related tasks (Section 6.4), keeping in mind the ethical concerns of communities affected by this research (Section 7).

## 2 Related Work

**Sexual Harassment Disclosures on Social Media** Several works have focused on identifying sexual violence (Leatherman, 2011), harassment and sexism (Wekerle et al., 2018; Manikonda et al., 2018b) in social media posts by analyzing factors such as linguistic themes, social engagement, and lexical attributes. Jha and Mamidi (2017) experimented with algorithms such as SVM and BiLSTM along with fastText to categorize hostility of sexist posts. (Parikh et al., 2019) proposed a multi-label CNN-based neural architecture along with word and sentence level embeddings for identifying variants of sexism present in online social platforms. Chowdhury et al. (2019) emphasized the use of linguistic themes, contextual meta-data, and semantic cues for evaluating human behaviors related to sex-

ual abuse disclosures. All of these works have dealt with modeling sexual disclosure narratives as single-task learning problems and were restricted to label specific detection (Marwa et al., 2018; Sawhney et al., 2020).

**Multitask Learning** Frameworks for learning representations across two different sources within the same domain follow multitask learning (Caruana, 1997). The ability to utilize knowledge from various sources compensates for missing data and complements existing meta-data (Tan et al., 2013; Ding et al., 2014), thus allowing for effective sharing of task-invariant features (Caruana, 1997; Zhang and Wang, 2016; Zhang et al., 2018). MTL has been utilized for name error recognition (Cheng et al., 2015), tagging-chunking (Collobert et al., 2011), machine translation (Luong et al., 2015) and relation extraction (Gupta et al., 2016). Liu et al. (2017) used shared and private latent features leveraging multitask learning for different text classification tasks. Rajamanickam et al. (2020); Duong et al. (2016); Liu et al. (2016) proposed a joint framework for modeling abuse and emotion detection and showed improvements over STL and transfer learning. Akhtar et al. (2018) proposed a multitask ensemble architecture for jointly modeling emotion, sentiment, and intensity, which gave improvements over single-label classification.

### 3 Problem Description

We aim to analyze different perspectives of the complex narratives pertaining to the #MeToo movement on social media platforms. Specifically, given a tweet text, we formulate for it a multi-label multi-class classification problem with definitions taken from previous works (ElSherief et al., 2018)

- *Stance Detection*: Determining the opinion of the author of a tweet, regarding a particular target of interest (Augenstein et al., 2016). Stance detection is categorized into three classes: *Support* for when the author favors the #MeToo movement or it’s cause; *Opposition*, representing opposing stance or indifference towards the movement; or *Neither*, when the text does not have a clear viewpoint (Mohammad and Turney, 2013).
- *Hate Speech Identification*: Detection of hate speech involves labeling the tweets as *Directed Hate* if the comment is targeted towards an individual or an entity, *Generalized Hate* if

it is targeted towards a community or a section of people or *Neither* otherwise (Basile et al., 2019).

- *Sarcasm Detection*: Given a tweet  $t_i$ , we aim to map it to either be *Sarcastic* or *Not Sarcastic* based on the presence of implicit sarcastic tone of the post (Bamman and Smith, 2015).
- *Dialogue Act Classification*: These are a function of a speaker’s utterance during a conversation, for example, question, answer, suggestion, etc., and are classified into three classes, namely *Allegation* (when the author intends to allege an individual or group of sexual misconduct) (Hutchings, 2012), *Justification* (tweets where the author is justifying their actions), and *Refutation* (for when the author refutes any accusation with or without evidence) (Gautam et al., 2020).

**Modeling Settings** To validate MTL’s performance across different domains, we also experiment with emotion detection as the *auxiliary* task. We aim to predict one or more of the several emotions representing the affective state of the authors - (*anger, disgust, anticipation, fear, joy, love, optimism, pessimism, sadness, surprise* and *trust*). We conceptualize three diverse problem settings and compare them to analyze MTL *within* and *across* domains. These are (i) *Single Task Learning*: Independent optimization of the four mentioned tasks associated with sexual abuse disclosure narrative classification, (ii) *Homogeneous Multitask Learning*: Simultaneous optimization of a pair selected from the four tasks associated with the sexual abuse disclosure posts, and (iii) *Heterogeneous Multitask Learning*: Classification of narratives associated with sexual abuse disclosure as the *primary* task and emotion detection as the *auxiliary* task.

## 4 Methodology

### 4.1 Text Encoding

Building on the success of transformer-based models in NLP, we chose BERTweet (Dat Quoc Nguyen and Nguyen, 2020), a pre-trained language model trained on 850 million English tweets. BERTweet has been trained with the same training procedure as RoBERTa (Liu et al., 2019) and has the same model configuration as the BERT base architecture (Devlin et al., 2019). The key component in

Task	Label	#Samples	Text
Relevance	Relevant	7,249	Guys are pissed off at [name] for affecting the credibility of a sexual assault survivor. Only men and r*p* enablers are questioning the movement in today's times. #Attack #BringTheChange.
	Stance	3,074	Thank you [name] for your courage passion and fight for #MeToo. It gives [name] strength to overcoming all this in front of people. Hope this inspires others as well to bring more stories. #Survivor .
Hate Speech	Directed	419	The progressive video by [user] shows the lady as a stripper, but are upset when [name] calls this movement bogus. Can't believe lies especially when so much has happened. #Fake #MeToo.
	Generalized	281	Life comes hard at [name]. Desperate for [name] approval, she tries to subvert the risk of her tawdry dating habits [URL] and disgusted company. Didn't even flinch once before saying this. #Fake.
Sarcasm	Sarcastic	220	These are not involved at all in this .... #MeToo has nothing with hating people. Its just a strategy by feminist h*e to get their a**es rich by manipulating people and asking for money. #Feminists
	Allegation	578	Thankfully the #HimToo movement will encourage [name] to put his d*c* inside his pant, out of the fear of getting publicly criticized by others. #MeToo. Is this really story of the decade, LOL !!! #SpeakOut
Dialogue Acts	Justification	292	Shut up now, [name], you said nothing when 10 women accused [name] of sexual harassment [URL] in the premises. Instead [name] remained silent among all this with eyes wide out. #IBelieveSurvivors [URL]
	Refutation	216	[name] embodies #MeToo movement, writing and spreading fake message loud and clear. Push false narrative and wrong ideology among the youth. This would get the job done. #BringOutTheTruth
			[name] says #MeToo is a trap, set by left wingers. Don't take the bait at all. The right shouldn't worry least, especially because of the involvement of [name]. This has happened far too many times in the past.

Table 1: Distribution of labels and examples for all tasks in #MeTooMA dataset. The tweets have been paraphrased for anonymity reasons and personally identifiable information has been censored. **We want to caution the readers that examples in this paper, though censored for profanity might contain offensive language.**

transformer-based models is the token level self-attention (Vaswani et al., 2017) that enables them to generate dynamic contextualized embeddings as opposed to static embeddings of GloVe (Pennington et al., 2014). Let  $(w_1, w_2, \dots, w_n)$  represent the sequence of tokens from a given tweet  $t$ . These tokens are pre-processed and passed through BERTweet<sup>3</sup>. We consider embeddings from the last layer of BERTweet and obtain an embedding  $e_i$  for a given tweet  $t_i$ . Embedding for each tweet is of dimension  $m \times k$ , where  $k$  represents the dimension size of BERT based model and  $m$  represents the maximum length for the tweets.

$$e_i = BERTweet(t_i) \quad (1)$$

These representations from Equation 1 are passed through a stacked BiLSTM encoder. Dropout is then applied to these encoded representations  $h^{(t)}$  (Equation 4 represents general formulation for both the tasks). These are then passed to a BiLSTM decoder, followed by a dropout layer and then a linear output layer to get output  $o^{(p)}$  ( $p$  representing primary task) or  $o^{(a)}$  ( $a$  representing auxiliary task).

$$\overrightarrow{h}_t^{(f)} = BiLSTM^{(f)}(e_t, h_{t-1}^{(f)}) \quad (2)$$

$$\overleftarrow{h}_t^{(b)} = BiLSTM^{(b)}(e_t, h_{t+1}^{(b)}) \quad (3)$$

$$h_t = [\overrightarrow{h}_t^{(f)}, \overleftarrow{h}_{T-t}^{(b)}] \quad (4)$$

## 4.2 Single Task Learning

We treat the task of categorizing narratives related to sexual abuse disclosure – *Stance*, *Hate Speech*,

*Sarcasm* and *Dialogue Acts*, independently. Each STL model is given an input representation  $e$  (Equation 1). Within the proposed tasks for classifying sexual abuse disclosure narrative for the tweets related to the #MeToo movement (Section 3), we use *sigmoid* activation for *Sarcasm* detection (whose classification outputs are binary) and *softmax* activation for all other tasks for the final output layer.

**Model Optimization** To account for the imbalance present among the labels, we use *class-balanced focal loss* as the optimization loss function (Cui et al., 2019), as formulated in Equation 5. Given a sample class  $i$  containing  $n_i$  samples in total, it adds a weighting factor of  $\frac{(1-\beta)}{(1-\beta^{n_i})}$  with parameters  $\beta \in [0,1)$ , where  $n_y$  is the number of samples in the ground truth class  $y$ . The proposed class-balanced term is model agnostic.  $p$  represents predicted class probabilities and  $\mathcal{L}$  represents the choice of the loss function (*binary cross entropy* for Sarcasm and *categorical cross entropy* for others).

$$CB(p, y) = \frac{1 - \beta}{1 - \beta^{n_y}} \mathcal{L}(p, y) \quad (5)$$

As for the multilabel emotion classification task, the unnormalized output (assuming one or more of 11 different emotions) is subjected to a *Sigmoid* activation, and the network is optimized using *binary cross-entropy* (BCE) as:

$$L_{BCE} = -\frac{1}{N} \sum_{i=1}^N y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i)) \quad (6)$$

where  $N$  is the number of training samples,  $y$  and  $p(y)$  denotes true and predicted labels respectively.

<sup>3</sup>Implementation used for BERTweet is available [here](#)

### 4.3 Multitask Learning

For our MTL approach, we use two optimization objectives: one for the *primary* task, which can be any of the proposed tasks for classifying tweets related to #MeToo movement (Section 3), and other for the *auxiliary* task, which can be either a task related to classifying sexual abuse disclosure for #MeToo movement (Homogeneous MTL) or emotion classification task (Heterogeneous MTL). The two objectives are weighted by a parameter  $\gamma$ , which controls the importance placed on the *auxiliary* task ( $1 - \gamma$  for the *primary* task).

Multitask learning frameworks are generally built using either of these two approaches: *hard parameter sharing* or *soft parameter sharing*. In a hard parameter sharing model (Caruana, 1997), both the primary and *auxiliary* tasks have a shared encoder followed by separate task-specific network branches, and the shared encoder is updated by both the tasks alternately. On the other hand, in the soft parameter sharing approach, tasks have different encoders with independent parameters, and the distance between their parameters is regularized using a regularization constraint (Duong et al., 2015; Yang and Hospedales, 2016), to encourage the parameters to be similar.

**Flexible Cross-Stitched Parameter Sharing Architecture:** We design our model so that the task-agnostic textual feature representations benefit from hard sharing while the regularization of the task-specific features can be learned according to task pair settings. We call our approach *flexible cross-stitched parameter sharing*, presented in Figure 2. Specifically, we train two separate models (one for each task) in tandem while also having a shared encoder that is updated by both of them and weighted joint learning of *primary* task decoder parameters that are tuned specifically for the task. This allows both the models to have their own set of parameters while also encouraging knowledge transfer via the shared encoder weights.

For each training pass of the *primary* task, the input representation  $e^{(p)}$  is passed through (a) *stacked BiLSTM encoder* and (b) *stacked shared BiLSTM encoder*. This results in two contextualized word representations  $(h_1^{(p)}, h_2^{(p)}, \dots, h_n^{(p)})$  and  $(h_1^{(s)}, h_2^{(s)}, \dots, h_n^{(s)})$ , where superscript ( $p$ ) is used to denote the representations resulting from encoder in the *primary* task model and superscript ( $s$ ) is used to denote the ones from shared encoder. We

calculate the weighted summation of these two representations -  $\tilde{h}^{(p)}$ , using two learnable parameters,  $\alpha^{(p)}$  and  $\alpha^{(s)}$  (where  $\alpha^{(p)} + \alpha^{(s)} = 1$ ), as formulated in Equation 7 to regulate the information resulting from the two encoders (Figure 2).

$$\tilde{h}^{(p)} = \alpha^{(p)}h^{(p)} + \alpha^{(s)}h^{(s)} \quad (7)$$

Such an approach to aggregate information flow from two encoders has facilitated success in prior Multitask learning settings as well (Rajamanickam et al., 2020; Dankers et al., 2019). As for our *auxiliary* task, we pass the embeddings  $e^{(a)}$  through only the shared encoder ( $h^{(a)} = h^{(s)}$ ), followed by a dropout layer. We use this architecture for Heterogeneous MTL experiments. For Homogeneous MTL ones, we employ **hard parameter sharing model** due to statistical out-performance in this scenario. This technique consists of a single stacked encoder that is shared and updated by both tasks related to identifying narratives related to sexual abuse disclosures within #MeToo movement, followed by task-specific branches. The shared representations from the encoder are passed through the dropout layer.

These output representations (in the case of both *Homogeneous* and *Heterogeneous* experiments) are passed through respective BiLSTM decoders and dropout layers to get the final representation  $m^{(p)}$  and  $m^{(a)}$ , respectively for both the tasks. The *auxiliary* network branch is optimized using either Equation 5 (Class Balanced Focal Loss) or Equation 6 (Binary Cross Entropy), depending upon whether the *auxiliary* task is associated with identifying sexual abuse disclosure narratives or emotions. These output representations  $m^{(p)}$  and  $m^{(a)}$  are passed through a linear output layer to get unnormalized outputs  $o^{(p)}$  and  $o^{(a)}$  respectively. *Sigmoid* activation function is used for *Sarcasm* detection and the *emotion classification task*, and *Softmax* activation for others.

## 5 Experiments

### 5.1 Data

MTL framework traditionally improves generalization by leveraging the domain-specific information due to the relatedness of the tasks present in the training signals (Caruana, 1997); hence we use two publicly available datasets mined from Twitter:

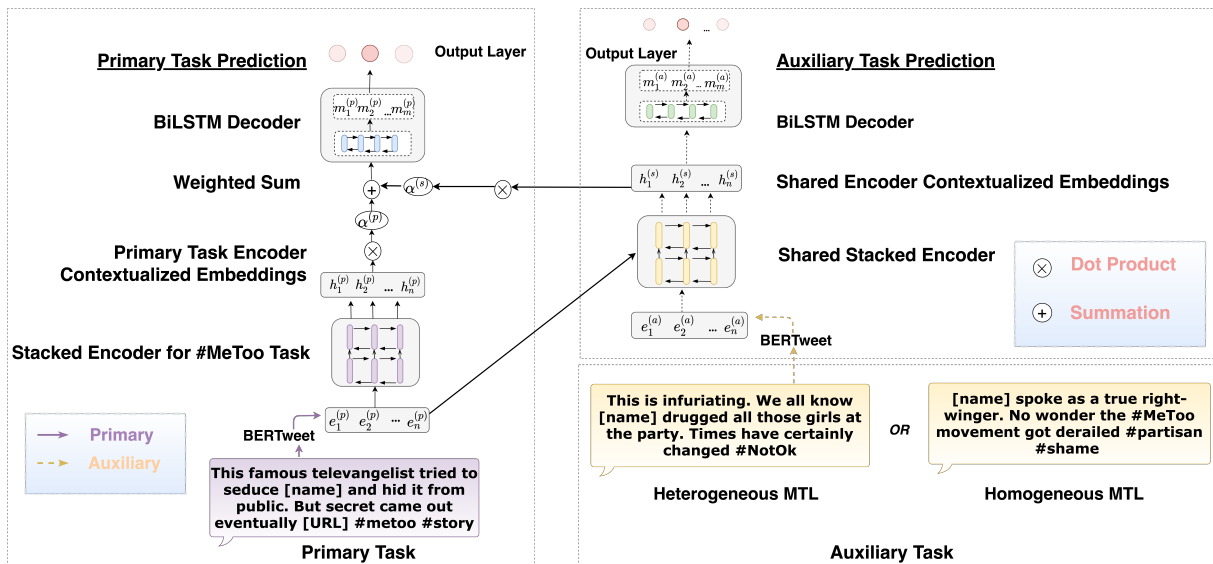


Figure 2: Flexible Cross-stitched Parameter Sharing Architecture. The embedding representations  $(e_1^{(p)}, e_2^{(p)}, \dots, e_n^{(p)})$  and  $(e_1^{(a)}, e_2^{(a)}, \dots, e_n^{(a)})$  identify BERTweet word-level embeddings for the *primary* and *auxiliary* task respectively. The different arrows are used to indicate the alternate passes of the *primary* task (solid arrows) and *auxiliary* task (dotted arrows). Two controllable parameters  $\alpha^{(p)}$  and  $\alpha^{(s)}$  are used to control information flow from task-specific and shared encoder respectively, for the *primary* task.

**Sexual Abuse Disclosures - #MeTooMA** This dataset<sup>4</sup> has 9,973 tweets and covers different mutually non-exclusive linguistic annotations related to the #MeToo movement (Gautam et al., 2020). The distribution and statistics about various labels are present in Table 1 and Section 3. We present an instance associated with each of the proposed tasks in Table 1. For our experiments, we focus only on tweets that are annotated as *relevant* to the #MeToo movement.

**Emotions - SemEval18** This dataset<sup>5</sup> has been taken from SemEval-2018 Task-1 (Mohammad et al., 2018) and covers emotion-specific labels representing the mental state of the authors of the tweets. It consists of 10,986 tweets distributed across 11 emotion labels – (*anger, disgust, anticipation, fear, joy, love, optimism, pessimism, sadness, surprise* and *trust*), each being a binary label to indicate the presence of a particular emotion.

## 5.2 Task Specific Setting

**Single Task Learning** STL experiments optimize each of the tasks associated with identifying narratives related to sexual abuse disclosures within #MeToo movement (Section 3) and emotion

detection, independently. We experiment with two distinct embedding spaces – GloVe-Twitter and BERTweet. Based on the superior performance of BERTweet with respect to GloVe-Twitter, we preferred it for further experimentation and studies.

**Homogeneous Multitask Learning** For this setup, we test the simultaneous optimization of two different tasks - both related to sexual harassment disclosure narratives, with one of them being *primary* and another coupled as the *auxiliary*. The results were obtained for a total of 12 pairs.

**Heterogeneous Multitask Learning** In these sets of experiments, we evaluate the positive transfer of representations *across* datasets by considering the identification of narratives associated with sexual abuse disclosure as the *primary* task and emotion detection as the *auxiliary* task.

## 5.3 Experimental Setup

**Preprocessing** We pre-process tweet text by (i) normalizing *user mentions* and *URLs*, and (ii) translating the emoticon into text (Hutto and Gilbert, 2014). For tokenization, we use *Tweet Tokenizer* from NLTK.<sup>6</sup>

<sup>4</sup>The publicly available dataset can be found at <https://doi.org/10.7910/DVN/JN4EYU>.

<sup>5</sup><https://competitions.codalab.org/competitions/17751>

<sup>6</sup><https://www.nltk.org/>

**Hyperparameters** For our model<sup>7</sup> hyperparameters were tuned on the validation set to find the best configurations. We use a pre-trained BERTweet model to extract 768-dimensional token-level embeddings. Grid search was performed to find the optimal value of hyperparameters and their range is summarized as: size of BiLSTM and dense layers  $\{128, 256, 512\}$ , embedding size  $d \in \{100, 200, 300\}$ , dropout  $\delta \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$ , learning rate  $\lambda \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ , weight decay  $\omega \in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$ , optimizer  $\{Adam, Adadelta\}$ , batch size  $b \in \{32, 64, 128\}$  and epochs ( $< 100$ ). For the MTL experiments, we tune the weightage of the *auxiliary* task ( $\gamma \in [0.1, 0.9]$  with intervals of 0.1) for each task pair.

For each task associated with identifying narratives pertaining to the #MeToo movement in the MTL setup, its value is considered as the one where the model performance improved the most and for both the tasks. For instance, we find the optimal value of  $\gamma$  for hate speech (as the *auxiliary* task) to be 0.4 in all Homogeneous task cases and of emotion detection to be 0.2 for the Heterogeneous tasks. For the MTL experiments,  $\alpha^p$  and  $\alpha^s$  are learnable and tuned on the validation loss. The encoders consist of two stacked BiLSTM’s with hidden size = 128. BiLSTM classifier has hidden size = 256, and the number of units in the penultimate dense layer is 128. Dropout is set to 0.3. For all our experiments, we use *Adam* optimizer (Kingma and Ba, 2014) and initialize model weights using *Xavier* initialization (Glorot and Bengio, 2010). We set the batch size to 128 and the learning rate to  $1e-3$ .

**Training** All models were trained until convergence for both *primary* and *auxiliary* tasks. For our MTL experiments, the training process involves alternating between *primary* and *auxiliary* task steps, with each task having its own loss function. All experiments are run using stratified 5-fold cross-validation. We report the average macro F1 scores across the 5 folds to account for imbalance, as previously used in multi-label settings (Zhang and Zhou, 2013).

<sup>7</sup>We used Keras with Tensorflow backend for implementing the models.

Task	ST	HS	SA	DI
ST	31.80	31.67	<b>32.41</b>	32.20
HS	<b>31.82</b>	31.78	31.64	31.80
SA	49.63	49.69	49.16	<b>49.79</b>
DI	<b>23.54</b>	23.42	23.20	23.41

Table 2: F1 macro score for pair-wise MTL (non-diagonal elements) and STL (diagonal elements - top left corner to bottom right corner). Rows denote the *primary* task and columns denote the *auxiliary* task in case of MTL. ST = Stance; HS = Hate Speech; SA = Sarcasm; DI = Dialogue. **Bold** denotes the highest score for that task.

Task	Homogeneous	Heterogeneous
Stance	32.41 $\pm$ 0.01 (SA)	<b>32.62 <math>\pm</math> 0.03</b>
Hate Speech	31.82 $\pm$ 0.02 (ST)	<b>32.01 <math>\pm</math> 0.01</b>
Sarcasm	<b>49.79 <math>\pm</math> 0.03</b> (DI)	49.50 $\pm$ 0.04
Dialogue	<b>23.54 <math>\pm</math> 0.03</b> (ST)	23.16 $\pm$ 0.06

Table 3: Best F1-scores obtained for Homogeneous MTL (Table 2) and Heterogeneous MTL experiments. Heterogeneous MTL experiments represent multitask learning performed with emotion identification as the *auxiliary* task. The best results highlighted in **bold**.

## 6 Results and Discussion

### 6.1 Single Task Learning

The aim of this paper is not limited to achieving the state of the art performance in terms of evaluation metrics but rather to conduct a thorough study to compare and contrast different methodologies for the benefit of the research community. As per our hypothesis and preliminary results on STL experiments on the #MeTooMA dataset, models trained using BERTweet embeddings perform far better than GloVe-Twitter. This is largely true because BERTweet is specifically pre-trained on English tweets and is better suited to handle Twitter-specific data, typically having a short length, informal grammar, and irregular vocabulary (e.g., abbreviations and typographical errors) (Kireyev et al., 2009).

### 6.2 Single Task Learning vis-a-vis Homogeneous Multitask Learning

Learning the affective states in the #MeTooMA dataset is challenging due to the inherently subjective nature of the tweets coupled with limitations on the data’s size. Multitask learning achieves significant performance gains in terms of macro F1 score, as shown in Table 2 for all task pairs. The diagonal results represented in **green** denote the **baseline** STL results whereas ones highlighted in shades of **blue** represent results for pair-wise

Tweet Text	STL	Homogeneous MTL	Heterogeneous MTL
$T_1$ – [name] says that nobody should be <b>ashamed</b> . Don't be scared and let it <b>bury</b> and <b>corrode</b> your soul. It gives <b>hope</b> through the <b>pain</b> , please visit [URL]. Lets speak up. #FightBack	Support	Support	Support
$T_2$ – Saying that #MeToo movement could save lives, would be a <b>grave</b> mistake, those individuals deserve it, such <b>swindling</b> . This is the proof of a disabled mindset among those guys. #Resist #RiseUp.	Neither	Support	Oppose
$T_3$ – When I ran my side production for the movie, people said that only <b>p*nsi*s</b> and <b>h*m*s</b> worked there because I had zero <b>tolerance</b> for sexual harassment in my unit. This is for all screeching people.	Neither	Gen Hate	Gen Hate
$T_4$ – I believe that this <b>ideology</b> must be broken! Society now has created <b>stigma</b> and its time we move <b>forward</b> to a new way of thinking for all girls. Please spread the word. #metoo #Healing.	Neither	Support	Support
$T_5$ – [name] says #MeToo is a <b>trap</b> , set by the left wingers. Don't take the <b>bait</b> . The right shouldn't worry least-bit, especially because of the involvement of [name] in these circumstances. #BringTheTruth	Neither	Refutation	Neither
$T_6$ – [name] embodies #MeToo movement today, writing and spreading fake messages <b>loud</b> and <b>clear</b> . Push <b>false</b> <b>narrative</b> and <b>wrong</b> <b>ideology</b> among the youth. See this [URL]. #Liar #YesAllWomen.	Neither	Justification	Neither

Table 4: Qualitative analysis of the performance obtained by MTL architecture on some samples. The color intensity of each word corresponds to the token-level attention score given by BERTweet. **Green** denotes correct prediction and **Yellow** denotes incorrect prediction. Tweets have been paraphrased to prevent user identification.

Homogeneous MTL with row identifying *primary* task and columns denoting *auxiliary* task. The higher performance of Homogeneous MTL can be inferred to be indicative of better generalization when pairs of tasks are jointly modeled. Interestingly, these tasks show their best performance with the selective counterparts in the Homogeneous MTL setup. *Stance* detection is strongly coupled with *Sarcasm* labeling, and the same is seen to be true for *Hate Speech* classification and *Stance* identification. This selective out-performance of specific pairs of tasks can be attributed to the high correlation between the tasks themselves (Frenda, 2018; Gautam et al., 2020). For instance, the offensive text is often strongly coupled with sarcasm, as *wit* is a common linguistic denominator for understanding the intended meaning of phrases related to anger (Badlani et al., 2019). We further detail this through examples in Section 6.4.

### 6.3 Heterogeneous Multitask Learning

Results in Table 3 demonstrate that the Heterogeneous MTL setup achieves higher performance than Homogeneous MTL under similar settings in two out of four task pairs<sup>8</sup> - *Stance* and *Hate Speech* detection by the margins of **+0.21** and **+0.19** respectively. For the other two tasks, the performance of Heterogeneous MTL is very close if not better than Homogeneous MTL. These findings are in line with the claim supporting the generalizability across tasks in the #MeTooMA dataset, which is

<sup>8</sup>We show only the best combinations of the Homogeneous task in the table for brevity.

highly correlated to emotion recognition. This is indicative of positive knowledge transfer between the two domains. Such joint optimization boosts the overall performance of both *primary* and *auxiliary* tasks through parameter sharing to learn common representations that may be mutually beneficial to both related tasks.

### 6.4 Qualitative Analysis

To emphasize our proposed approach, we perform a qualitative study by handpicking examples from the dataset. We analyze token-level attention assigned to individual terms by BERTweet, where color intensity corresponds to the attention score. These results are shown in Table 4. We infer that Homogeneous and Heterogeneous multitask learning shows superior performance in every instance compared to STL. Learning effective features across the joint formulation of pair-wise tasks in Homogeneous MTL is evident from  $T_4$ , where BERT’s self-attention allots a higher weight to words such as *ideology*, *stigma*, and *forward* in line with the actual label as *Support*.

Similarly for  $T_5$ , highlighted terms such as *trap* and *bait* are indicative of the opposing nature of the tweets, hence identified as belonging to *Refutation*. On the other hand, due to positive knowledge transfer from the emotion recognition task, Heterogeneous MTL obtains better performance in several cases. Words such as *grave*, *mistake* and *swindling* in  $T_2$  connoted a negative emotion, hence accordingly being identified as belonging to the *Oppose* category. Similarly, terms such as



*hope* and *pain* were given higher token-level attention in  $T_1$  emphasizing a positive emotion and thus can be correlated with belonging to the *Support* category. An interesting observation is the presence of named entities in  $T_5$  and  $T_6$ , resulting in the incorrect prediction via Heterogeneous MTL. Therefore, a limitation of the single task learning and Heterogeneous MTL is the inability to mitigate the effect of named entities or specific events in the text to influence the knowledge transfer and create negative shared representations.

## 7 Ethical Concerns and Discussion

Analyzing social media data of individuals discussing sexual harassment disclosures and exploitation in public spheres necessitates the need to safeguard the ethics and privacy of individuals (Tusinski Berg, 2019). We address these:

**Generalization** We acknowledge that the limitations of the experiments might get amplified due to the highly subjective nature of this challenging problem. Therefore it would not be fair to conduct a population-centric analysis based on inferences from this work.

**Confidentiality** Individual consent was not sought from social media users as the data was publicly available. Disclosure of sexual harassment information on public forums may have been met with public backlash and apathy. Therefore the social reputation of the *accuser* and the *accused* would be at a peril (McDonald, 2019). Hence, the authors were aware not to make any automated interventions, as any attempts to contact individuals could be seen as personally intrusive and might also reveal their social information (Fiesler and Proferes, 2018).

**Bias & Discrimination** Social support discussions on social media platforms gave victims the liberty to describe their instances of sexual exploitation and abuse (Manikonda et al., 2018a). The authors are aware of the potential inevitable sampling biases that may be present in the data. Importance has to be placed on mitigating the bias against certain minority groups, which might get amplified due to the sensitive nature of social discussions (Hellwig and Sinno, 2017).

## 8 Conclusion

In this work, we have proposed a *flexible cross-stitched multitask learning framework* for the de-

tection of narratives linked with sexual abuse disclosure on social media. Our methodology takes advantage of the affective features from emotions and related tasks to encourage knowledge transfer and attain *auxiliary* knowledge. Qualitative and quantitative results demonstrate how joint optimization of *Stance* detection and *Sarcasm* identification benefit each other, indicating their relatedness and dependence on each other. Similarly, we observe that tasks like *Hate-Speech* classification and *Stance* labeling benefit from each other and from emotion detection, thus reinforcing the benefit of joint linguistic learning between the related tasks. In the future, we aim to explore how this joint learning paradigm can be effectively leveraged for improving performance on downstream tasks like emotion analysis, identifying suicidal tendencies among abuse survivors. Application from this work also has utility for problems such as identification of patterns of reported sexual harassment narratives, hate speech detection, the spread of rumors and fake news, and entity extraction for digital vigilantism (Yuce et al., 2014; Hosterman et al., 2018).

## Acknowledgments

Rajiv Ratn Shah is partly supported by the Infosys Center for AI and Center for Design and New Media at IIIT Delhi.

## References

- Md Shad Akhtar, Deepanway Ghosal, Asif Ekbal, Pushpak Bhattacharyya, and Sadao Kurohashi. 2018. A multi-task ensemble framework for emotion, sentiment and intensity prediction. *arXiv preprint arXiv:1808.01216*.
- Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer.
- Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. *Stance detection with bidirectional conditional encoding*. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas. Association for Computational Linguistics.
- Rohan Badlani, Nishit Asnani, and Manan Rai. 2019. Disambiguating sentiment: An ensemble of humour, sarcasm, and hate speech features for sentiment classification. *W-NUT 2019*, page 337.

- David Bamman and Noah A Smith. 2015. Contextualized sarcasm detection on twitter. In *Ninth International AAAI Conference on Web and Social Media*.
- Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 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*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.
- Hao Cheng, Hao Fang, and Mari Ostendorf. 2015. Open-domain name error detection using a multi-task rnn. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 737–746.
- Arijit Ghosh Chowdhury, Ramit Sawhney, Rajiv Shah, and Debanjan Mahata. 2019. # youtoo? detection of personal recollections of sexual harassment on social media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2527–2537.
- Rosemary Clark-Parsons. 2019. “i see you, i believe you, i stand with you”:# metoo and the performance of networked feminist visibility. *Feminist Media Studies*, pages 1–19.
- Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537.
- Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9268–9277.
- Verna Dankers, Marek Rei, Martha Lewis, and Ekaterina Shutova. 2019. [Modelling the interplay of metaphor and emotion through multitask learning](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2218–2229, Hong Kong, China. Association for Computational Linguistics.
- Thanh Vu Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. *arXiv preprint*, arXiv:2005.10200.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- Zhengming Ding, Shao Ming, and Yun Fu. 2014. Latent low-rank transfer subspace learning for missing modality recognition. In *Twenty-eighth AAAI conference on artificial intelligence*.
- Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. [Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 845–850, Beijing, China. Association for Computational Linguistics.
- Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning crosslingual word embeddings without bilingual corpora. *arXiv preprint arXiv:1606.09403*.
- Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. *arXiv preprint arXiv:1804.04257*.
- Casey Fiesler and Nicholas Proferes. 2018. “participant” perceptions of twitter research ethics. *Social Media+ Society*, 4(1):2056305118763366.
- Mary Anne Franks. 2019. Witch hunts: Free speech,# metoo, and the fear of women’s words. *U. Chi. Legal F.*, page 123.
- Simona Frenda. 2018. The role of sarcasm in hate speech. a multilingual perspective. In *e Doctoral Symposium of the XXXIV International Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*, pages 13–17. Lloret, E.; Saquete, E.; Mart inez-Barco, P.; Moreno, I.
- Akash Gautam, Puneet Mathur, Rakesh Gosangi, Debanjan Mahata, Ramit Sawhney, and Rajiv Ratn Shah. 2020. # metooma: Multi-aspect annotations of tweets related to the metoo movement. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 209–216.
- Xavier Glorot and Y. Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. *Journal of Machine Learning Research - Proceedings Track*, 9:249–256.
- Pankaj Gupta, Hinrich Schütze, and Bernt Andrassy. 2016. Table filling multi-task recurrent neural network for joint entity and relation extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2537–2547.

- Timothy Hellwig and Abdulkader Sinno. 2017. Different groups, different threats: public attitudes towards immigrants. *Journal of Ethnic and Migration Studies*, 43(3):339–358.
- Alec R Hosterman, Naomi R Johnson, Ryan Stouffer, and Steven Herring. 2018. Twitter, social support messages, and the #metoo movement. *The Journal of Social Media in Society*, 7(2):69–91.
- Chris Hutchings. 2012. Commercial use of facebook and twitter—risks and rewards. *Computer Fraud & Security*, 2012(6):19–20.
- Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.
- Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the second workshop on NLP and computational social science*, pages 7–16.
- Sweta Karlekar and Mohit Bansal. 2018. Safecity: Understanding diverse forms of sexual harassment personal stories. *arXiv preprint arXiv:1809.04739*.
- Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.
- Kirill Kireyev, Leysia Palen, and Kenneth Anderson. 2009. Applications of topics models to analysis of disaster-related twitter data. In *NIPS workshop on applications for topic models: text and beyond*, volume 1. Canada: Whistler.
- Linda Lane and Ulla-Carin Hedin. 2020. The promise of the #metoo movement for preventing and reporting sexual harassment. *Available at SSRN*.
- Janie Leatherman. 2011. *Sexual violence and armed conflict*. Polity.
- Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*.
- Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. *arXiv preprint arXiv:1704.05742*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#).
- Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.
- Lydia Manikonda, Ghazaleh Beigi, Subbarao Kambhampati, and Huan Liu. 2018a. #metoo through the lens of social media. In *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation*, pages 104–110. Springer.
- Lydia Manikonda, Ghazaleh Beigi, Huan Liu, and Subbarao Kambhampati. 2018b. Twitter for sparking a movement, reddit for sharing the moment: #metoo through the lens of social media. *arXiv preprint arXiv:1803.08022*.
- Tolba Marwa, Ouadfel Salima, and Meshoul Souham. 2018. Deep learning for online harassment detection in tweets. In *2018 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, pages 1–5. IEEE.
- Aubri F McDonald. 2019. Framing #metoo: Assessing the power and unintended consequences of a social media movement to address sexual assault. In *Handbook of Sexual Assault and Sexual Assault Prevention*, pages 79–107. Springer.
- Brad McKenna and Hameed Chughtai. 2020. Resistance and sexuality in virtual worlds: An lgbt perspective. *Computers in Human Behavior*, 105:106199.
- Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.
- Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. 29(3):436–465.
- Pulkit Parikh, Harika Abburi, Pinkesh Badjatiya, Radhika Krishnan, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2019. Multi-label categorization of accounts of sexism using a neural framework. *arXiv preprint arXiv:1910.04602*.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. [GloVe: Global vectors for word representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Santhosh Rajamanickam, Pushkar Mishra, Helen Yanakoudakis, and Ekaterina Shutova. 2020. Joint modelling of emotion and abusive language detection. *arXiv preprint arXiv:2005.14028*.
- Mannila Sandhu, C Danielle Vinson, Vijay K Mago, and Philippe J Giabbanelli. 2019. From associations to sarcasm: Mining the shift of opinions regarding the supreme court on twitter. *Online Social Networks and Media*, 14:100054.

- Ramit Sawhney, Akash Kumar Gautam, and Rajiv Ratn Shah. 2020. Bmgc 2020 grand challenge: Multi-aspect analysis of the metoo movement on twitter. In *2020 IEEE Sixth International Conference on Multi-media Big Data (BigMM)*, pages 481–484. IEEE.
- Sima Sharifirad, Borna Jafarpour, and Stan Matwin. 2018. Boosting text classification performance on sexist tweets by text augmentation and text generation using a combination of knowledge graphs. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pages 107–114.
- Ashwini Tambe. 2018. Reckoning with the silences of# metoo. *Feminist Studies*, 44(1):197–203.
- Ben Tan, Erheng Zhong, Evan Wei Xiang, and Qiang Yang. 2013. Multi-transfer: Transfer learning with multiple views and multiple sources. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 243–251. SIAM.
- Kati Tusinski Berg. 2019. The ethics of exploring gender issues in a time of# metoo. *Journal of Media Ethics*, 34(1):52–56.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinukas Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.
- Christine Wekerle, Negar Vakili, Sherry H Stewart, and Tara Black. 2018. The utility of twitter as a tool for increasing reach of research on sexual violence. *Child abuse & neglect*, 85:220–228.
- Yongxin Yang and Timothy M. Hospedales. 2016. [Trace norm regularised deep multi-task learning](#).
- Serpil T Yuce, Nitin Agarwal, Rolf T Wigand, Merylna Lim, and Rebecca S Robinson. 2014. Bridging women rights networks: Analyzing interconnected online collective actions. *Journal of Global Information Management (JGIM)*, 22(4):1–20.
- Min-Ling Zhang and Zhi-Hua Zhou. 2013. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837.
- Xiaodong Zhang and Houfeng Wang. 2016. A joint model of intent determination and slot filling for spoken language understanding. In *IJCAI*, volume 16, pages 2993–2999.
- Yuxiang Zhang, Jiamei Fu, Dongyu She, Ying Zhang, Senzhang Wang, and Jufeng Yang. 2018. Text emotion distribution learning via multi-task convolutional neural network. In *IJCAI*, pages 4595–4601.