# Uncover Sexual Harassment Patterns from Personal Stories by Joint Key Element Extraction and Categorization

**Yingchi Liu[1], Quanzhi Li[1], Marika Cifor[2], Xiaozhong Liu[3], Qiong Zhang[1] and Luo Si[1]**
[1]Alibaba Group, US
[2]University of Washington, USA
[3]Indiana University Bloomington, USA
{yingchi.liu, quanzhi.li, qz.zhang,luo.si}@alibaba-inc.com
mcifor@uw.edu, liu237@indiana.edu

## Abstract

The number of personal stories about sexual harassment shared online has increased exponentially in recent years. This is in part inspired by the #MeToo and #TimesUp movements. Safecity is an online forum for people who experienced or witnessed sexual harassment to share their personal experiences. It has collected >10,000 stories so far. Sexual harassment occurred in a variety of situations, and categorization of the stories and extraction of their key elements will provide great help for the related parties to understand and address sexual harassment. In this study, we manually annotated those stories with labels in the dimensions of location, time, and harassers' characteristics, and marked the key elements related to these dimensions. Furthermore, we applied natural language processing technologies with joint learning schemes to automatically categorize these stories in those dimensions and extract key elements at the same time. We also uncovered significant patterns from the categorized sexual harassment stories. We believe our annotated data set, proposed algorithms, and analysis will help people who have been harassed, authorities, researchers and other related parties in various ways, such as automatically filling reports, enlightening the public in order to prevent future harassment, and enabling more effective, faster action to be taken.

## 1 Introduction

Sexual violence, including harassment, is a pervasive, worldwide problem with a long history. This global problem has finally become a mainstream issue thanks to the efforts of survivors and advocates. Statistics show that girls and women are put at high risk of experiencing harassment. Women have about a 3 in 5 chance of experiencing sexual harassment, whereas men have slightly less than 1 in 5 chance (Quinnipiac, 2017; EEOC, 2018;

Goldstein, 2018). While women in developing countries are facing distinct challenges with sexual violence (Lea et al., 2017), however sexual violence is ubiquitous. In the United States, for example, there are on average >300,000 people who are sexually assaulted every year (Morgan and Truman, 2018). Additionally, these numbers could be underestimated, due to reasons like guilt, blame, doubt and fear, which stopped many survivors from reporting (Griffith, 2018). Social media can be a more open and accessible channel for those who have experienced harassment to be empowered to freely share their traumatic experiences and to raise awareness of the vast scale of sexual harassment, which then allows us to understand and actively address abusive behavior as part of larger efforts to prevent future sexual harassment. The deadly gang rape of a medical student on a Delhi bus in 2012 was a catalyst for protest and action, including the development of Safecity, which uses online and mobile technology to work towards ending sexual harassment and assault. More recently, the #MeToo and #TimesUp movements, further demonstrate how reporting personal stories on social media can raise awareness and empower women. Millions of people around the world have come forward and shared their stories. Instead of being bystanders, more and more people become up-standers, who take action to protest against sexual harassment online. The stories of people who experienced harassment can be studied to identify different patterns of sexual harassment, which can enable solutions to be developed to make streets safer and to keep women and girls more secure when navigating city spaces (Karlekar and Bansal, 2018). In this paper, we demonstrated the application of natural language processing (NLP) technologies to uncover harassment patterns from social media data. We made three key contributions:

2328

1. Safecity[1] is the largest publicly-available online forum for reporting sexual harassment (Karlekar and Bansal, 2018). We annotated about 10,000 personal stories from Safecity with the key elements, including information of harasser (i.e. the words describing the harasser), time, location and the trigger words (i.e. the phrases indicate the harassment that occurred). The key elements are important for studying the patterns of harassment and victimology (Griffith, 2018; Ceccato, 2017). Furthermore, we also associated each story with five labels that characterize the story in multiple dimensions (i.e. age of harasser, single/multiple harasser(s), type of harasser, type of location and time of day). The annotation data are available online.[2]

2. We proposed joint learning NLP models that use convolutional neural network (CNN) (Lecun and Bengio, 1995) and bi-directional long short-term memory (BiLSTM) (Schuster and Paliwal, 1997; Hochreiter and Schmidhuber, 1997) as basic units. Our models can automatically extract the key elements from the sexual harassment stories and at the same time categorize the stories in different dimensions. The proposed models outperformed the single task models, and achieved higher than previously reported accuracy in classifications of harassment forms (Karlekar and Bansal, 2018).

3. We uncovered significant patterns from the categorized sexual harassment stories.

## 2 Related Work

Conventional surveys and reports are often used to study sexual harassment, but harassment on these is usually under-reported (Goldstein, 2018; Griffith, 2018). The high volume of social media data available online can provide us a much larger collection of firsthand stories of sexual harassment. Social media data has already been used to analyze and predict distinct societal and health issues, in order to improve the understanding of wide-reaching societal concerns, including mental health, detecting domestic abuse, and cyberbullying (Balani and De Choudhury, 2015; Schrading et al., 2015; Ziegele et al., 2018; Agrawal and Awekar, 2018).

There are a very limited number of studies on

---

Figure 1: An example of the annotation.

sexual harassment stories shared online. Karlekar and Bansal (2018) were the first group to our knowledge that applied NLP to analyze large amount ( ∼10,000) of sexual harassment stories. Although their CNN-RNN classification models demonstrated high performance on classifying the forms of harassment, only the top 3 majority forms were studied. In order to study the details of the sexual harassment, the trigger words are crucial. Additionally, research indicated that both situational factors and person (or individual difference) factors contribute to sexual harassment (Hitlan et al., 2009). Therefore, the information about perpetrators needs to be extracted as well as the location and time of events. Karlekar and Bansal (2018) applied several visualization techniques in order to capture such information, but it was not obtained explicitly. Our preliminary research demonstrated automatic extraction of key element and story classification in separate steps (Liu et al., 2019). In this paper, we proposed joint learning NLP models to directly extract the information of the harasser, time, location and trigger word as key elements and categorize the harassment stories in five dimensions as well. Our approach can provide an avenue to automatically uncover nuanced circumstances informing sexual harassment from online stories.

## 3 Data Collection and Annotation

We obtained 9,892 stories of sexual harassment incidents that was reported on Safecity. Those stories include a text description, along with tags of the forms of harassment, e.g. commenting, ogling and groping. A dataset of these stories was published by Karlekar and Bansal (2018). In addition to the forms of harassment, we manually annotated each story with the key elements (i.e. "harasser", "time", "location", "trigger"), because they are essential to uncover the harassment patterns. An example is shown in Figure 1. Furthermore, we also assigned each story classification labels in five dimensions (Table 1). The detailed definitions of classifications in all dimensions are explained below.

**Age of Harasser**: Individual difference such as age can affect harassment behaviors. Therefore,

| Dimensions | Classes/Groups |
|---|---|
| Age of Harasser | 0 - unspecified, 1 - teenagers or young people, 2 - adults |
| Single/Multiple Harasser(s) | 0 - unspecified, 1 - single harasser, 2 - multiple harassers |
| Type of Harasser | 0 - unspecified, 1 - relative, 2 - teacher, 3 - classmate, 4 - friend, 5 - neighbor, 6 - conductor/driver, 7 - work-related, 8 - police/guard, 9 - other |
| Type of Location | 0 - unspecified, 1 - street, 2 - transportation, 3 - station/stop, 4 - private places, 5 - shopping places, 6 - neighborhood, 7 - park, 8 - hotel, 9 - bush/woods, 10 - parking lot, 11 - in/near school, 12 - restaurant, 13 - other |
| Time of Day | 0 - unspecified, 1 - day (5am to 6pm), 2 - evening or night |

Table 1: Definition of classes in different dimensions about sexual harassment.

we studied the harassers in two age groups, young and adult. Young people in this paper refer to people in the early 20s or younger.

**Single/Multiple Harasser(s)**: Harassers may behave differently in groups than they do alone.

**Type of Harasser**: Person factors in harassment include the common relationships or titles of the harassers. Additionally, the reactions of people who experience harassment may vary with the harassers' relations to themselves (Griffith, 2018). We defined 10 groups with respects to the harassers' relationships or titles. We put conductors and drivers in one group, as they both work on the public transportation. Police and guards are put in the same category, because they are employed to provide security. Manager, supervisors, and colleagues are in the work-related group. The others are described by their names.

**Type of Location**: It will be helpful to reveal the places where harassment most frequently occurs (Ceccato, 2017; Karlekar and Bansal, 2018). We defined 14 types of locations. "Station/stop" refers to places where people wait for public transportation or buy tickets. Private places include survivors' or harassers' home, places of parties and etc. The others are described by their names.

**Time of Day**: The time of an incident may be reported as *"in evening"* or at a specific time, e.g. *"10 pm"*. We considered that 5 am to 6 pm as day time, and the rest of the day as the night.

Because many of the stories collected are short, many do not contain all of the key elements. For example, *"A man came near to her tried to be physical with her ."*. The time and location are unknown from the story. In addition, the harassers were strangers to those they harassed in many cases. For instance, *"My friend was standing in the queue to pay bill and was ogled by a group of boys."*, we can only learn that there were multiple young harassers, but the type of harasser is

unclear. The missing information is hence marked as "unspecified". It is different from the label "other", which means the information is provided but the number of them is too small to be represented by a group, for example, a "trader".

All the data were labeled by two annotators with training. Inter-rater agreement was measured by Cohen's kappa coefficient, ranging from 0.71 to 0.91 for classifications in different dimensions and 0.75 for key element extraction (details can refer to Table 1 in supplementary file). The disagreements were reviewed by a third annotator and a final decision was made.

## 4 Proposed Models

The key elements can be very informative when categorizing the incidents. For instance, in Figure 1, with identified key elements, one can easily categorize the incident in dimensions of "age of harasser" (adult), "single/multiple harasser(s)" (single), "type of harasser" (unspecified), "type of location" (park) , "time of day" (day time). Therefore, we proposed two joint learning schemes to extract the key elements and categorize the incidents together. In the models' names, "J", "A", "SA" stand for joint learning, attention, and supervised attention, respectively.

### 4.1 CNN Based Joint Learning Models

In Figure 2, the first proposed structure consists of two layers of CNN modules.

**J-CNN:** To predict the type of key element, it is essential for the CNN model to capture the context information around each word. Therefore, the word along with its surrounding context of a fixed window size was converted into a context sequence. Assuming a window size of $2l + 1$ around the target word $w_0$, the context sequence is $[(w_{-l}, w_{-l+1}, ...w_0, ...w_{l-1}, w_l)]$, where $w_i (i \in [-l, l])$ stands for the $ith$ word from $w_0$.
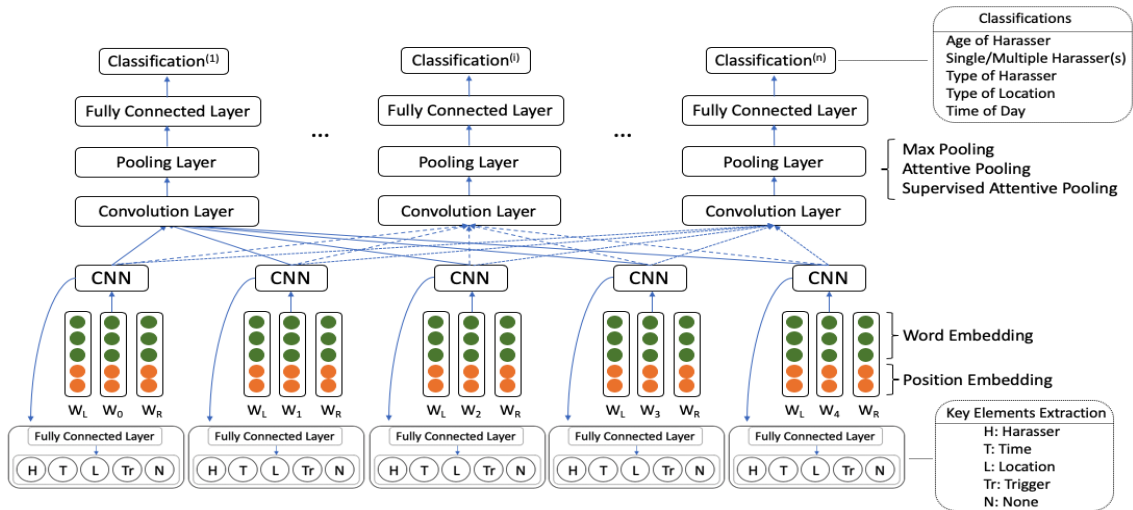
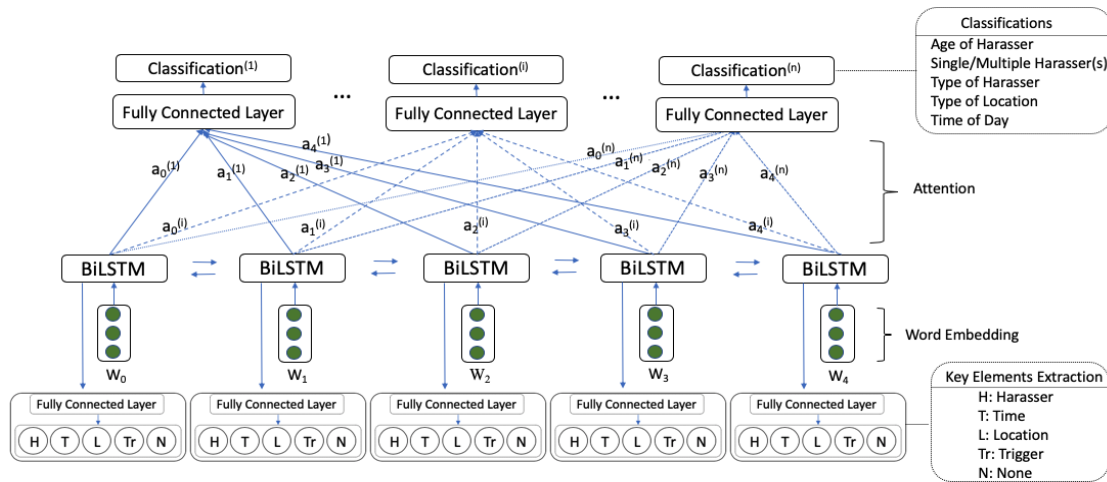Figure 2: CNN based Joint learning Model. $W_L$ and $W_R$ are the left and right context around each word.



Figure 3: BiLSM based Joint Learning Model. Here we use an input of five words as an example.

Because the context of the two consecutive words in the original text are only off by one position, it will be difficult for the CNN model to detect the difference. Therefore, the position of each word in this context sequence is crucial information for the CNN model to make the correct predictions (Nguyen and Grishman, 2015). That position was embedded as a $p$ dimensional vector, where $p$ is a hyperparameter. The position embeddings were learned at the training stage. Each word in the original text was then converted into a sequence of the concatenation of word and position embeddings. Such sequence was fed into the CNN modules in the first layer of the model, which output the high level word representation ($h_i, i \in [0, n-1]$, where n is the number of input words). The high level word representation was then passed into a fully connected layer, to pre-

dict the key element type for the word. The CNN modules in this layer share the same parameters.

We input the sequence of high level word representations ($h_i$) from the first layer into another layer of multiple CNN modules to categorize the harassment incident in each dimension (Figure 2). Inside each CNN module, the sequence of word representations were first passed through a convolution layer to generate a sequence of new feature vectors ($C = [c_0, c_1, ...c_q]$). This vector sequence ($C$) was then fed into a max pooling layer. This is followed by a fully connected layer. Modules in this layer do not share parameters across classification tasks.

**J-ACNN:** We also experimented with attentive pooling, by replacing the max pooling layer. The attention layer aggregates the sequence of feature vectors ($C$) by measuring the contribution of each

vector to form the high level representation of the harassment story. Specifically,

$$u_i = tanh(W_\omega c_i + b_\omega) \tag{1}$$

$$\alpha_i = \frac{exp(u_i^T u_\omega)}{\sum_i exp(u_i^T u_w)} \tag{2}$$

$$v = \sum_i \alpha_i c_i \tag{3}$$

That is, a fully connected layer with non-linear activation was applied to each vector $c_i$ to get its hidden representation $u_i$. The similarity of $u_i$ with a context vector $u_w$ was measured and get normalized through a softmax function, as the importance weight $\alpha_i$. The final representation of the incident story $v$ was an aggregation of all the feature vectors weighted by $\alpha_i$. $W_\omega$, $b_\omega$ and $u_w$ were learned during training.

The final representation ($v$) was passed into one fully connected layer for each classification task. We also applied different attention layers for different classifications, because the classification modules categorize the incident in different dimensions, their focuses vary. For example, to classify "time of day", one needs to focus on the time phrases, but pays more attention to harassers when classifying "age of harasser".

**J-SACNN:** To further exploit the information of the key elements, we applied supervision (Zhao et al., 2018) to the attentive pooling layer, with the annotated key element types of the words as ground truth. For instance, in classification of "age of harasser", the ground truth attention labels for words with key element types of "harasser" are 1 and others are 0. To conform to the CNN structure, we applied convolution to the sequence of ground truth attention labels, with the same window size ($w$) that was applied to the word sequence (Eq. 4).

$$\alpha^*(t) = W \circ [e_t : e_{t+w-1}] \tag{4}$$

where $\circ$ is element-wise multiplication, $e_t$ is the ground truth attention label, and the $W \in R^{w \times 1}$ is a constant matrix with all elements equal to 1. $\alpha^*$ was normalized through a softmax function and used as ground truth weight values of the vector sequence ($C$) output from the convolution layer. The loss was calculated between learned attention $\alpha$ and $\alpha^*$ (Eq. 5), and added to the total loss.

$$E(\alpha, \alpha^*) = \sum_{i=0}^{q} (\alpha_i - \alpha_i^*)^2 \tag{5}$$

## 4.2 BiLSTM Based Joint Learning Models

**J-BiLSTM:** The model input the sequence of word embeddings to the BiLSTM layer. To extract key elements, the hidden states from the forward and backward LSTM cells were concatenated and used as word representations to predict the key element types.

To classify the harassment story in different dimensions, concatenation of the forward and backward final states of BiLSTM layer was used as document level representation of the story.

**J-ABiLSTM:** We also experimented on BiLSTM model with the attention layer to aggregate the outputs from BiLSTM layer (Figure 3). The aggregation of the outputs was used as document level representation.

**J-SABiLSTM:** Similarly, we experimented with the supervised attention.

In all the models, softmax function was used to calculate the probabilities at the prediction step, and the cross entropy losses from extraction and classification tasks were added together. In case of supervised attention, the loss defined in Eq. 5 was added to the total loss as well. We applied the stochastic gradient descent algorithm with mini-batches and the AdaDelta update Rule (rho=0.95 and epsilon=1e-6) (Zeiler, 2012; Feng et al., 2016). The gradients were computed using back-propagation. During training, we also optimized the word and position embeddings.

## 5 Experiments and Results

### 5.1 Experimental Settings

**Data Splits**: We used the same splits of train, develop, and test sets used by Karlekar and Bansal (Karlekar and Bansal, 2018), with 7201, 990 and 1701 stories, respectively. In this study, we only considered single label classifications.

**Baseline Models**: CNN and BiLSTM models that perform classification and extraction separately were used as baseline models. In classification, we also experimented with BiLSTM with the attention layer. To demonstrate that the improvement came from joint learning structure rather the two layer structure in J-CNN, we investigated the same model structure without training on key element extraction. We use J-CNN* to denote it.

**Preprocess**: All the texts were converted to lowercase and preprocessed by removing non-alphanumeric characters, excluding ". ! ? " . The word embeddings were pre-trained using fastText

(Bojanowski et al., 2016) with dimension equaling 100.

**Hyperparameters**: For the CNN model, the filter size was chosen to be (1,2,3,4), with 50 filters per filter size. Batch size was set to 50 and the dropout rate was 0.5. The BiLSTM model comprises two layers of one directional LSTM. Every LSTM cell has 50 hidden units. The dropout rate was 0.25. Attention size was 50.

## 5.2 Results and Discussions

We compared joint learning models with the single task models. Results are averages from five experiments. Although not much improvement was achieved in key element extraction (Figure 2), classification performance improved significantly with joint learning schemes (Table 3). Significance t-test results are shown in Table 2 in the supplementary file.

**BiLSTM Based Models:** Joint learning BiLSTM with attention outperformed single task BiLSTM models. One reason is that it directed the attention of the model to the correct part of the text. For example,

**S1:** *"when i was returning my home after finishing my class . i was in queue to get on the micro bus and there was a girl opposite to me just then a young man tried to touch her on the breast ."*

**S2:** *"when i was returning my home after finishing my class . i was in queue to get on the micro bus and there was a girl opposite to me just then a young man tried to touch her on the breast ."*

**S3:** *"when i was returning my home after finishing my class . i was in queue to get on the micro bus and there was a girl opposite to me just then a young man tried to touch her on the breast ."*

In S1, the regular BiLSTM with attention model for classification on "age of harasser" put some attention on phrases other than the harasser, and hence aggregated noise. This could explain why the regular BiLSTM model got lower performance than the CNN model. However, when training with key element extractions, it put almost all attention on the harasser "young man" (S2), which helped the model make correct prediction of "young harasser". When predicting the "type of location" (S3), the joint learning model directed its attention to "micro bus".

**CNN Based Models:** Since CNN is efficient for capturing the most useful information (Chen et al., 2015), it is quite suitable for the classifica-

|  | Accuracy | Macro F1 |
| --- | --- | --- |
| BiLSTM | 92.1 | 79.9 |
| CNN | **92.6** | 79.5 |
| J-BiLSTM | 92.1 | 79.2 |
| J-ABiLSTM | 92.0 | 79.7 |
| J-SABiLSTM | 92.0 | 79.1 |
| J-CNN | 92.5 | 80.1 |
| J-ACNN | 92.4 | **80.4** |
| J-SACNN | 92.4 | 80.1 |

Table 2: Key element extraction results.

|  |  | Age | Single/ Multiple | Type of Harasser | Loc | Time |
| --- | --- | --- | --- | --- | --- | --- |
| Accuracy | BiLSTM | 90.7 | 91.1 | 91.0 | 80.3 | 97.0 |
|  | ABiLSTM | 90.0 | 91.8 | 91.4 | 81.6 | 97.0 |
|  | CNN | 91.6 | 92.8 | 93.2 | 83.6 | 97.4 |
|  | J-CNN* | 91.5 | 92.7 | 92.3 | 82.6 | 97.2 |
|  | J-BiLSTM | 90.8 | 91.8 | 91.3 | 78.3 | 94.8 |
|  | J-ABiLSTM | 92.4 | **94.0** | 92.3 | **85.1** | 97.7 |
|  | J-SABiLSTM | 92.4 | 93.7 | 92.1 | 84.8 | 97.4 |
|  | J-CNN | 92.5 | 93.8 | **93.3** | 84.2 | 98.0 |
|  | J-ACNN | **92.8** | 93.3 | 93.1 | 84.2 | 97.9 |
|  | J-SACNN | 92.5 | 93.8 | 92.7 | 83.1 | **98.0** |
| Macro F1 | BiLSTM | 90.4 | 90.3 | 38.6 | 44.3 | 93.8 |
|  | ABiLSTM | 89.7 | 91.0 | 33.5 | 45.5 | 93.6 |
|  | CNN | 91.5 | 92.1 | 46.7 | 48.1 | 94.4 |
|  | J-CNN* | 91.4 | 92.0 | 46.8 | 45.1 | 94.5 |
|  | J-BiLSTM | 90.4 | 91.1 | 34.1 | 34.1 | 87.7 |
|  | J-ABiLSTM | 92.4 | **93.4** | 37.9 | 48.5 | 95.1 |
|  | J-SABiLSTM | 92.4 | 93.0 | 37.8 | 48.5 | 94.4 |
|  | J-CNN | 92.4 | 93.1 | 49.5 | 48.6 | **95.6** |
|  | J-ACNN | **92.8** | 92.8 | 47.7 | 48.5 | 95.3 |
|  | J-SACNN | 92.4 | **93.4** | **52.8** | **50.5** | 95.2 |

Table 3: Classification accuracy and macro F1 of the models. The best scores are in bold.

tion tasks in this study. It achieved better performance than the BiLSTM model. The joint learning method boosted the performance even higher. This is because the classifications are related to the extracted key elements, and the word representation learned by the first layer of CNNs (Figure 2) is more informative than word embedding. By plotting of t-SNEs (Maaten and Hinton, 2008) of the two kinds of word vectors, we can see the word representations in the joint learning model made the words more separable (Figure 1 in supplementary file). In addition, no improvement was found with the J-CNN* model, which demonstrated the joint learning with extraction is essential for the improvement.

With supervised attentive pooling, the model can get additional knowledge from key element labels. It helped the model in cases when certain location phrases were mentioned but the incidents did not happen at those locations. For instance, *"I*

| Models | commenting | ogling | groping |
|---|---|---|---|
| CNN* | 80.9 | 82.2 | 86.0 |
| BiLSTM* | 81.0 | 82.2 | 86.2 |
| CNN+RNN* | 81.6 | 84.1 | 86.5 |
| J-BiLSTM | 81.7 | 83.3 | 87.1 |
| J-ABiLSTM | 82.1 | 83.2 | 87.9 |
| J-SABiLSTM | 82.4 | 83.1 | 87.9 |
| J-CNN | 81.9 | **84.4** | 87.4 |
| J-ACNN | **82.6** | 83.8 | 88.1 |
| J-SACNN | 82.3 | 83.6 | **88.7** |

Table 4: Harassment form classification accuracy of models. * Reported by Karlekar and Bansal (2018)

*was followed on my way home .”*, max pooling will very likely to predict it as "private places". But, it is actually unknown. In other cases, with supervised attentive pooling, the model can distinguish "metro" and "metro station", which are "transportation" and "stop/station" respectively. Therefore, the model further improved on classifications on "type of location" with supervised attention in terms of macro F1. For some tasks, like "time of day", there are fewer cases with such disambiguation and hence max pooling worked well. Supervised attention improved macro F1 in location and harasser classifications, because it made more correct predictions in cases that mentioned location and harasser. But the majority did not mention them. Therefore, the accuracy of J-SACNN did not increase, compared with the other models.

**Classification on Harassment Forms:** In Table 4, we also compared the performance of binary classifications on harassment forms with the results reported by Karlekar and Bansal (2018). Joint learning models achieved higher accuracy. In some harassment stories, the whole text or a span of the text consists of trigger words of multiple forms, such as *"stare, whistles, start to sing, commenting"*. The supervised attention mechanism will force the model to look at all such words rather than just the one related to the harassment form for classification and hence it can introduce noise. This can explain why J-SACNN got lower accuracy in two of the harassment form classifications, compared to J-ACNN. In addition, J-CNN model did best in "ogling" classification.

# 6 Patterns of Sexual Harassment

We plotted the distribution of harassment incidents in each categorization dimension (Figure 4). It displays statistics that provide important evidence as to the scale of harassment and that can serve as the
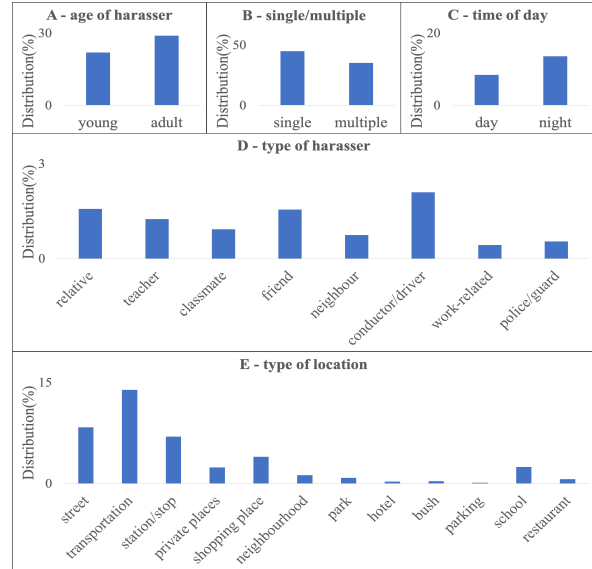


Figure 4: Distributions of incidents. A) Distributions over age of harasser, B) over single/multiple harasser(s), C) over time of day, D) over type of harasser. E) over type of location.

basis for more effective interventions to be developed by authorities ranging from advocacy organizations to policy makers. It provides evidence to support some commonly assumed factors about harassment: First, we demonstrate that harassment occurred more frequently during the night time than the day time. Second, it shows that besides unspecified strangers (not shown in the figure), conductors and drivers are top the list of identified types of harassers, followed by friends and relatives.

Furthermore, we uncovered that there exist strong correlations between the age of perpetrators and the location of harassment, between the single/multiple harasser(s) and location, and between age and single/multiple harasser(s) (Figure 5). The significance of the correlation is tested by chi-square independence with $p$ value less than 0.05. Identifying these patterns will enable interventions to be differentiated for and targeted at specific populations. For instance, the young harassers often engage in harassment activities as groups. This points to the influence of peer pressure and masculine behavioral norms for men and boys on these activities. We also found that the majority of young perpetrators engaged in harassment behaviors on the streets. These findings suggest that interventions with young men and boys, who are readily influenced by peers, might be most effective when education is done peer-to-peer. It also points to the locations where such efforts
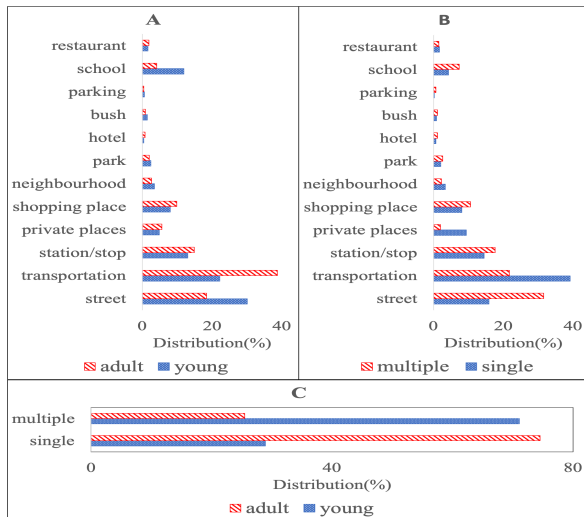
Figure 5: Distributions of incidents over two dimensions. A) Distributions of incidents A) with young/adult harassers at each location, B) with single/multiple harasser(s) at each location, C) across young/adult harassers and single/multiple harasser(s)
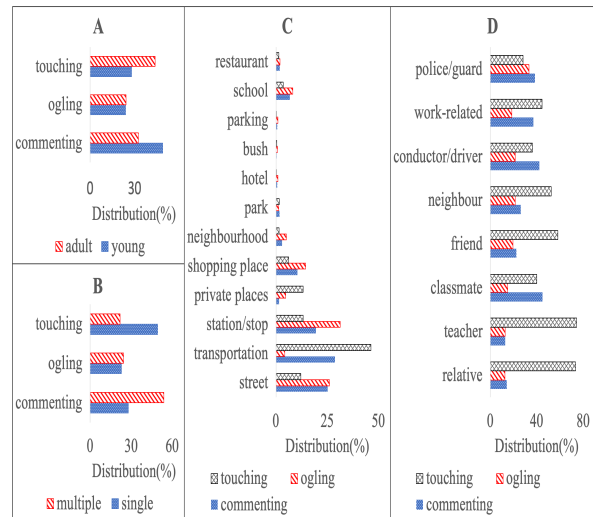


Figure 6: Distributions of incidents with harassment forms and different dimensions. Distributions of harassment forms A) within each age group, B) within single/multiple harasser(s), C) over locations, D) within each harasser type.

could be made, including both in schools and on the streets. In contrast, we found that adult perpetrators of sexual harassment are more likely to act alone. Most of the adult harassers engaged in harassment on public transportation. These differences in adult harassment activities and locations, mean that interventions should be responsive to these factors. For example, increasing the security measures on transit at key times and locations.

In addition, we also found that the correlations between the forms of harassment with the age, single/multiple harasser, type of harasser, and location (Figure 6). For example, young harassers are more likely to engage in behaviors of verbal harassment, rather than physical harassment as compared to adults. It was a single perpetrator that engaged in touching or groping more often, rather than groups of perpetrators. In contrast, commenting happened more frequently when harassers were in groups. Last but not least, public transportation is where people got indecently touched most frequently both by fellow passengers and by conductors and drivers. The nature and location of the harassment are particularly significant in developing strategies for those who are harassed or who witness the harassment to respond and manage the everyday threat of harassment. For example, some strategies will work best on public transport, a particular closed, shared space setting, while other strategies might be more effective on the open space of the street.

These results can provide valuable information for all members of the public. Sharing stories of harassment has been found by researchers to shift people's cognitive and emotional orientation towards their traumatic experiences (Dimond et al., 2013). Greater awareness of patterns and scale of harassment experiences promises to ensure those who have been subjected to this violence that they are not alone, empowering others to report incidents, and ensuring them that efforts are being made to prevent others from experiencing the same harassment. These results also provide various authorities tools to identify potential harassment patterns and to make more effective interventions to prevent further harassment incidents. For instance, the authorities can increase targeted educational efforts at youth and adults, and be guided in utilizing limited resources the most effectively to offer more safety measures, including policing and community-based responses. For example, focusing efforts on highly populated public transportation during the nighttime, when harassment is found to be most likely to occur.

## 7 Conclusions

We provided a large number of annotated personal stories of sexual harassment. Analyzing and identifying the social patterns of harassment behavior is essential to changing these patterns and social tolerance for them. We demonstrated the joint learning NLP models with strong performances to

automatically extract key elements and categorize the stories. Potentiality, the approaches and models proposed in this study can be applied to sexual harassment stories from other sources, which can process and summarize the harassment stories and help those who have experienced harassment and authorities to work faster, such as by automatically filing reports (Karlekar and Bansal, 2018). Furthermore, we discovered meaningful patterns in the situations where harassment commonly occurred. The volume of social media data is huge, and the more we can extract from these data, the more powerful we can be as part of the efforts to build a safer and more inclusive communities. Our work can increase the understanding of sexual harassment in society, ease the processing of such incidents by advocates and officials, and most importantly, raise awareness of this urgent problem.

## References

Sweta Agrawal and Amit Awekar. 2018. Deep learning for detecting cyberbullying across multiple social media platforms. *CoRR*, abs/1801.06482.

Sairam Balani and Munmun De Choudhury. 2015. Detecting and characterizing mental health related self-disclosure in social media. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. ACM.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Vania Ceccato. 2017. Women's victimisation and safety in transit environments. *Crime Prevention and Community Safety*, 19(3):163–167.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the ACL and the 7th International JCNLP*, pages 167–176.

Jill Dimond, Michaelanne Dye, Daphne Larose, and Amy S. Bruckman. 2013. Hollaback!: the role of storytelling online in a social movement organization. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, pages 477–490.

EEOC. 2018. Charges alleging sexual harassment.

Xiaocheng Feng, Lifu Huang, Duyu Tang, Heng Ji, Bing Qin, and Ting Liu. 2016. A language-independent neural network for event detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 66–71. ACL.

Allison Goldstein. 2018. Sexual harassment statistics: Is the truth in the numbers?

Linda Lewis Griffith. 2018. Guilt, doubt, fear: Why women don't report sexual assault.

Robert T. Hitlan, John B. Pryor, Matthew S. Hesson-McInnis, and Michael Olson. 2009. Antecedents of gender harassment: An analysis of person and situation factors. *Sex Roles*, 61(11):794–807.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Sweta Karlekar and Mohit Bansal. 2018. SafeCity: Understanding diverse forms of sexual harassment personal stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2805–2811. ACL.

Suzanne Goodney Lea, Elsa D'Silva, and Abhijith Asok. 2017. Women's strategies addressing sexual harassment and assault on public buses: an analysis of crowdsourced data. *Crime Prevention and Community Safety*, 19(3):227–239.

Yann Lecun and Yoshua Bengio. 1995. *Convolutional Networks for Images, Speech and Time Series*, pages 255–258. MIT Press.

Yingchi Liu, Quanzhi Li, Xiaongzhong Liu, Qiong Zhang, and Luo Si. 2019. Sexual harassment story classification and key information identification. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM'19)*.

L.J.P.V.D. Maaten and GE Hinton. 2008. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Rachel E. Morgan and Jennifer L. Truman. 2018. *Criminal Victimization, 2017*. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

Thien Huu Nguyen and Ralph Grishman. 2015. Event detection and domain adaptation with convolutional neural networks. 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 365–371. ACL.

Quinnipiac. 2017. 60% of u.s. women say they've been sexually harassed quinnipiac university national poll finds.

Nicolas Schrading, Cecilia Ovesdotter Alm, Ray Ptucha, and Christopher Homan. 2015. An analysis of domestic abuse discourse on reddit. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2577–2583. ACL.

M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *Trans. Sig. Proc.*, 45(11):2673–2681.

Matthew D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701.

Yue Zhao, Xiaolong Jin, Yuanzhuo Wang, and Xueqi Cheng. 2018. Document embedding enhanced event detection with hierarchical and supervised attention. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 414–419. ACL.

Marc Ziegele, Johannes Daxenberger, Oliver Quiring, and Iryna Gurevych. 2018. Developing automated measures to predict incivility in public online discussions on the facebook sites of established news media. In *Proceedings of the 68th Annual Conference of the International Communication Association (ICA)*.