Speak Up, Fight Back!
Detection of Social Media Disclosures of Sexual Harassment

Arijit Ghosh Chowdhury*
Manipal Institute of Technology, MAHE
arijit10@gmail.com

Ramit Sawhney*
Netaji Subhas Institute of Technology
ramits.co@nsit.net.in

Puneet Mathur
MIDAS, IIIT-Delhi
pmathur3k6@gmail.com

Debanjan Mahata
Bloomberg
dmahata@bloomberg.net

Rajiv Ratn Shah
MIDAS, IIIT-Delhi
rajivratn@iiitd.ac.in

Abstract

The #MeToo movement is an ongoing prevalent phenomenon on social media aiming to demonstrate the frequency and widespread of sexual harassment by providing a platform to speak up and narrate personal experiences of such harassment. The aggregation and analysis of such disclosures pave the way to the development of technology-based prevention of sexual harassment. We contend that the lack of specificity in generic sentence classification models may not be the best way to tackle text subtleties that intrinsically prevail in a classification task as complex as identifying disclosures of sexual harassment. We propose the Disclosure Language Model, a three-part ULMFiT architecture, consisting of a Language model, a Medium-Specific (Twitter) model, and a Task-Specific classifier to tackle this problem and create a manually annotated real-world dataset to test our technique on this, to show that using a Discourse Language Model often yields better classification performance over (i) Generic deep learning based sentence classification models (ii) existing models that rely on handcrafted stylistic features. An extensive comparison with state-of-the-art generic and specific models along with a detailed error analysis presents the case for our proposed methodology.

1 Introduction

Thirty-five percent of women, including people in the LGBTQIA+ community, are globally subjected to sexual or physical assault, according to a study by UN Women [1]. With the advent of the #MeToo movement (Lee, 2018), discussions about sexual abuse have finally seen the light as compared to before, without the fear of shame or retaliation. Abuse in general and sexual harassment, in particular, is one topic that is socially stigmatized and difficult for people to talk about in both non-computer-mediated and computer-mediated contexts. The Disclosure Processes Model (DPM) (Andalibi et al., 2016) examines when and why interpersonal disclosure may be beneficial and focuses on people with concealable stigmatized identities (e.g., abuse, rape) in non-computer-mediated contexts. It has been found that disclosure of abuse has positive psychological impacts (Manikonda et al., 2016); (McClain and Amar, 2013)), and the #MeToo movement has managed to make social media avenues like Twitter a safer place to share personal experiences.

The information gathered from these kinds of online discussions can be leveraged to create better campaigns for social change by analyzing how users react to these stories and obtaining a better insight into the consequences of sexual abuse. Prior studies noted that developing an automated framework for classifying a tweet is quite challenging due to the inherent complexity of the natural language constructs (Badjatiya et al., 2017).

Tweets are entirely different from other text forms like movie reviews and news forums. Tweets are often short and ambiguous because of the limitation of characters. There are more mis-

* Denotes equal contribution.

[1] http://www.unwomen.org/en/what-we-do/ending-violence-against-women/facts-and-figures

Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 136–146
Minneapolis, Minnesota, June 3 - 5, 2019.

spelled words, slangs, and acronyms on Twitter because of its casual form (Mahata et al., 2015). This motivates our study to build a medium-specific Language Model for the segregation of tweets containing disclosures of sexual harassment.

While there is a developing body of literature on the topic of identifying patterns in the language used on social media that analyze sexual harassment disclosure (Manikonda et al., 2018); (Andalibi et al., 2016), very few attempts have been made to segregate texts containing discussions about sexual abuse from texts containing personal recollections of sexual harassment experiences. Efforts have been made to segregate domestic abuse stories from Reddit by Schrading et al. (2015) and Karlekar and Bansal. However, these approaches do not take into consideration the model's domain understanding of the syntactic and semantic attributes of the specific medium in which the text is present.

In that regard, our paper makes two significant contributions.

1. Generation of a labeled real-world dataset for identifying social media disclosures of sexual abuse, by manual annotation.

2. Comparison of the proposed Medium-Specific Disclosure Language Model architecture for segregation of tweets containing disclosure, with various deep learning architectures and machine learning models, in terms of four evaluation metrics.

2 Related Work

Twitter is fast becoming the most widely used source for social media research, both in academia and in industry (Meghawat et al., 2018) (Shah and Zimmermann, 2017). Wekerle et al. (2018) have shown that Twitter is being used for increasing research on sexual violence. Using social media could support at-risk youth, professionals, and academics given the many strengths of employing such a knowledge mobilization tool. Previously, Twitter has been used to tackle mental health issues (Sawhney et al., 2018b) (Sawhney et al., 2018a) and for other social issues like detection of hate speech content online (Mathur et al., 2018). Mahata et al. (2018) have mobilized Twitter to detect information regarding personal intake of medicines. Social media use is free, easy to implement, available to difficult to access populations (e.g., victims of sexual violence), and can reduce the gap between research and practice. Bogen et al. (2018) discusses the social reactions to disclosures of sexual victimization on Twitter. This work suggests that online forums may offer a unique context for disclosing violence and receiving support. Khatua et al. (2018) have explored deep learning techniques to classify tweets of sexual violence, but have not explicitly focused on building a robust system that can detect recollections of personal stories of abuse.

Schrading et al. (2015) created the Reddit Domestic Abuse Dataset, to facilitate classification of domestic abuse stories using a combination of SVM and N-grams. Karlekar and Bansal improved upon this by using CNN-LSTMs, due to the complementary strengths of both these architectures. Reddit allows lengthy submissions, unlike Twitter, and therefore the use of standard English is more common. This allows natural language processing tools trained on standard English to function better. Our method explores the merits of using a Twitter-specific Language Model which can counter the shortcomings of using pre-trained word embeddings derived from other tasks, on a medium like Twitter where the language is informal, and the grammar is often ambiguous.

N-gram based Twitter Language Models (Vo et al., 2015) have been previously used to detect events and for analyzing Twitter conversations (Ritter et al., 2010). Atefeh and Khreich (2015) used Emoticon Smoothed Language Models for Twitter Sentiment Analysis. Rother and Rettberg (2018) used the ULMFiT model proposed by Howard and Ruder (2018) to detect offensive tweets in German. Manikonda et al. (2018) try to investigate social media posts discussing sexual abuse by analyzing factors such as *linguistic themes*, *social engagement*, and *emotional attributes*. Their work proves that Twitter is an effective source for human behavior analysis, based on several linguistic markers. Andalibi et al. (2016) attempt to characterize abuse related disclosures into different categories, based on different themes, like gender, support seeking nature, etc. Our study aims to bridge the gap between gathering information and analyzing social media disclosures of sexual abuse. Our approach suggests that the language used on Twitter can be treated as a separate language construct, with its own rules and restrictions that need to be ad-

dressed to capture subtle nuances and understand the context better.

3 Data

3.1 Data Collection

Typically, it has been difficult to extract data related to sexual harassment due to social stigma but now, an increasing number of people are turning to the Internet to vent their frustration, seek help and discuss sexual harassment issues. To maintain the privacy of the individuals in the dataset, we do not present direct quotes from any data, nor any identifying information.

Anonymized data was collected from microblogging website Twitter - specifically, content containing self-disclosures of sexual abuse from November 2016 to December 2018.

The creation of a new dataset mandates specific linguistic markers needed to be identified. Instead of developing a word list to represent this language, a corpus of words and phrases were developed using anonymized data from known Sexual Harassment forums [2] [3] [4].

User posts containing tags of *metoo*, *sexual violence* and *sexual harassment* were also collected from microblogging sites like Tumblr and Reddit. For e.g., subreddits like *r/traumatoolbox, r/rapecounseling, and r/survivorsofabuse*. Then the TF-IDF method was applied to these texts to determine words and phrases (1-grams, 2-grams and 3-grams) which frequently appeared in posts related to sexual harassment and violence. Finally, human annotators were asked to remove terms from this which were not based on sexual harassment, as well as duplicate terms. This process generated 70 words/phrases which were used as a basis for extraction of tweets.

The public Streaming API was used for the collection and extraction of recent and historical tweets. These texts were collected without knowing the sentiment or context. For example, when collecting tweets on the hashtag #metoo, it is not known initially whether the tweet has been posted for sexual assault awareness and prevention, or if the person is talking about their own experience of sexual abuse, or if the tweet reports an incident or a news report.

[2] http://www.aftersilence.org/
[3] https://pandys.org/
[4] http://isurvive.org/

was assaulted	*molested me*
raped me	*touched me*
groped	*I was stalked*
forced me	*#WhyIStayed*
#WhenIwas	*#NotOkay*
abusive	*relationship*
drugged	*underage*
inappropriate	*followed*
boyfriend	*workplace*

Table 1: Words/Phrases linked with Sexual Harassment

3.2 Data Annotation

Then, text posts equaling 5117 in all were collected which were subsequently human annotated. The annotators included Clinical Psychologists and Academia of Gender Studies. All the annotators had to review the entire dataset. The tweets were segregated based on the following criteria.

Is the user recollecting their personal experience of sexual harassment?

Every post was scrutinized and carefully analyzed by three independent annotators $H1$, $H2$ and $H3$ due to the subjectivity of text annotation. Ambiguous posts were set to the default level of *Non-Disclosure*.

The following annotation guidelines were followed.

- The default category for all posts is Non-Disclosure.
- The text is marked as Disclosure if it explicitly mentions a personal abuse experience; e.g., *"I was molested by my ex-boyfriend"*

  e.g., *"I was told by my boss that my skirt was too distracting."*
- Posts which mentioned other people's recollections were not marked as Disclosure; e.g.*"My friend's boss harassed her"*
- If the tone of the text is flippant. e.g.*"I can't play CS I got raped out there hahaha"*, then it is marked as Non-Disclosure
- Posts related to sexual harassment related news reports or incidents, e.g., *"Woman gang-raped by 12 men in Uttar Pradesh"*, are marked as Non-Disclosure.

- Posts about sexual harassment awareness e.g.*"Sexual assault and harassment are unfortunately issues that continue to plague our college community."*, are marked as Non-Disclosure.

Finally, after an agreement between the annotators (Table 4), 1126 tweets in the dataset (22% of the dataset) were annotated as Self-Disclosure with an average value of Cohen Kappas inter-annotator agreement κ = 0.83, while the rest fell into the category of Non-Disclosure. The imbalance of the dataset is encouraged to represent a realistic picture usually seen on social media websites. Our dataset is made publicly available [5], following the guidelines mentioned in Section 7 to facilitate further research and analysis on this very pertinent issue.

4 Methodology

4.1 Preprocessing

The following preprocessing steps were taken as a part of noise reduction: Extra white spaces, newlines, and special characters were removed from the sentences. All stopwords were removed. Stopwords corpus was taken from NLTK and was used to eliminate words which provide little to no information about individual tweets. URLs, screen names, hashtags(#), digits (0-9), and all Non-English words were removed from the dataset.

4.2 The Disclosure Language Model (DLM)

Previous studies show that traditional learning methods such as manual feature extraction or using representation learning methods followed by a linear classifier have been inefficient in comparison to recent deep learning methods (Khatua et al., 2018). Bag-of-words approaches tend to have a high recall but lead to high rates of false positives because lexical detection methods classify all messages containing particular terms only. Following this stream of research, our work considers deep learning techniques for the detection of social media disclosures of sexual harassment.

CNNs also have been able to generate state of the art results in text classification because of their ability to extract features from word embeddings (Kim, 2014). Recent approaches that concatenate embeddings derived from other tasks with the input at different layers (Maas et al. (2011)) still train from scratch and treat pre-trained embeddings as fixed parameters, limiting their usefulness.

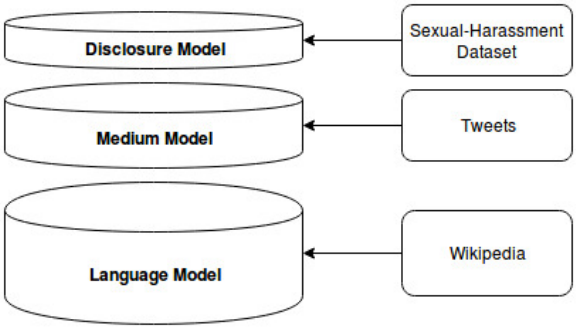

Figure 1: The Disclosure Language Model Overview

We propose a three-part Disclosure Classification method, based on the Universal Language Model Fine-tuning (ULMFiT) architecture, introduced by (Howard and Ruder, 2018) that enables robust inductive transfer learning for any NLP task, akin to fine-tuning ImageNet models: We use the 3-layer AWD-LSTM architecture proposed by Merity et al. (2017) using the same hyperparameters and no additions other than tuned dropout hyperparameters. Dropouts have been successful in feed-forward and convolutional neural networks, but applying dropouts similarly to an RNNs hidden state is ineffective as it disrupts the RNNs ability to retain long-term dependencies, and may cause overfitting. Our proposed method makes use of DropConnect (Merity et al., 2017), in which, instead of activations, a randomly selected subset of weights within the network is set to zero. Each unit thus receives input from a random subset of units in the previous layer. By performing dropout on the hidden-to-hidden weight matrices, overfitting can be prevented on the recurrent connections of the LSTM.

4.3 Classification

For every tweet $t_i \in D$, in the dataset, a binary valued value variable y_i is used, which can either be 0 or 1. The value 0 indicates that the text belongs to the Non-Disclosure category while 1 indicates Disclosure.

The training has been split into three parts as shown in $Figure 1$.

- **Language Model (LM)** - This model is trained from a large corpus of unlabeled data. In this case, a pre-trained Wikipedia Language Model was used.

[5] github.com/ramitsawhney27/NAACLSRW19meToo

Disclosure
 # WhenIWas 15 I was molested by my best friend
I was sexually assaulted by my step brother in 2009.
At 8 years old, an aldult family member sexually assaulted me.
I was 7 the first time I was sexually assaulted.
I was sexually assaulted by at least 3 different babysitters by the time I was 6 years old.

Table 2: Human Annotation examples for Self Disclosure.

Non-Disclosure
Sexual assault and harassment are unfortunately issues that continue to plague our community.
Trying to silence sexual assault victims is another one. The list goes on and on
Then call for people that cover up sexual assault like Jim Jordan to resign???
sexual assault on public transport is real
agreed! metoo is not just exclusively for women!

Table 3: Human Annotation examples for Non Disclosure

	H1	H2	H3
H1	–	0.74	0.88
H2	0.74	–	0.86
H3	0.88	0.86	–

Table 4: Cohen's Kappa for Annotators H_1, H_2, and H_3

Layer	Howard Dropout
Input	0.25
General	0.1
LSTM Internal	0.2
Embedding	0.02
Between LSTM Layers	0.15

Table 5: Dropout used by Howard and Ruder (2018)

- **Medium Model (MM)** - The Language Model is used as the basis to train a Medium Model (MM) from unlabeled data that matches the desired medium of the task (e.g., forum posts, newspaper articles or tweets). In our study the weights of the pre-trained Language Model are slowly re-trained on a subset of the Twitter Sentiment140 dataset [6]. This augmented vocabulary improves the model's domain understanding of Tweet syntax and semantics.

- **Disclosure Model (DM)** - Finally, a binary classifier is trained on top of the Medium Model from a labeled dataset. This approach facilitates the reuse of pre-trained models for the lower layers.

[6]https://www.kaggle.com/Jazzanova/sentiment140

5 Experiment Setup

5.1 Baselines

To make a fair comparison between all the models mentioned above, the experiments are conducted with respect to specific baselines.

Schrading et al. (2015) proposed the Domestic Abuse Disclosure (DAD) Model using the 1, 2, and 3-grams in the text, the predicates, and the semantic role labels as features, including TF-IDF and Bag of Words.

Andalibi et al. (2016) used a Self-Disclosure Analysis (SDA) Logistic Regression model with added features like TF-IDF and Char-N-grams, to characterize abuse-related disclosures by analyzing word occurrences in the texts.

In the experiments, we also evaluate and compare our model with several widely used baseline methods, mentioned in Table 6.

A small subset (10%) of the dataset is held back for testing on unseen data.

5.2 DLM Architectures and Parameters

Our method uses the Weight Dropped AWD-LSTM architecture used by , using the same hyperparameters and no additions other than tuned dropout hyperparameters. Embedding size is 400, the number of hidden activations per layer is 1150, and the number of layers used is 3. Two linear blocks with batch normalization and dropout have

Architecture	Specification
RNN (Liu et al., 2016)	Can efficiently represent more complex patterns than the shallow neural networks.
LSTM (Wang et al., 2018)	LSTMs are able to capture the long-term dependency among words in short texts.
Bi-LSTM (Tai et al., 2015)	At each time step, the hidden state of the Bidirectional LSTM is the concatenation of the forward and backward hidden states.
GRU (Rana, 2016):	Simplified variation of the LSTM. Combines the forget and input gates into a single update gate.
CNN (Kim, 2014)	Utilize layers with convolving filters that are applied to local features.
Very Deep-CNN (Conneau et al., 2016)	Operate directly at the character level and use only small convolutions and pooling operations.
Char-CNN (Zhang et al., 2015)	The model can understand abnormal character combinations and new languages.
fastText Bag of Tricks (Joulin et al., 2017)	Word features are averaged together to form sentence representations.
HATT (Yang et al., 2016):	Two levels of attention mechanisms applied at the word-and sentence-level.
DP CNN (Johnson and Zhang, 2017)	Low-complexity word-level CNN that can detect long associations.
R-CNN (Lai et al., 2015)	Uses a recurrent structure to capture contextual information when learning word representations.
CNN-LSTM (Zhou et al., 2015)	Utilizes CNN to extract higher-level phrase representations, which are fed into an LSTM.
A-CNN-LSTM (Yuan et al., 2018)	Combined C-LSTM model with additional attention mechanisms.
openAI-Transformer (Vaswani et al., 2017)	Generative pre-training and discriminative fine-tuning of a language model for a specific task.

Table 6: Baseline Specifications

Model	**Medium**	Type	
Language Model	Wikipedia	1,000,000,000	Unlabeled
Medium Model	Twitter	100,000	Unlabeled
Disclosure Model	Twitter	5117	Labeled

Table 7: Training Data Overview

been added to the model, with rectified linear unit activations for the intermediate layer and a softmax activation at the last layer.

The models use different configurations for back-propagation through time (BPTT), learning rate (LR), weight decay (WD), dropouts, cyclical learning rates (CLR) (Smith (2017)) and slanted triangular learning rates (STLR) (Howard and Ruder (2018)). Additionally, gradient clipping (Pascanu et al. (2013) has been applied to some of the models. The RNN hidden-to-hidden matrix uses a weight dropout for all the models. We train the models for 15 epochs.

For the CLR the four parameters are maximum

Architecture	Accuracy	Precision	Recall	F1
DAD Model	0.91	0.90	0.91	0.90
SDA Model	0.90	0.87	0.90	0.88
Word-CNN	0.92	0.68	0.95	0.79
LSTM	0.92	0.70	0.98	0.81
RNN	0.93	0.86	0.95	0.90
GRU	0.87	0.47	0.80	0.59
VD-CNN	0.93	0.89	0.90	0.89
CL-CNN	0.92	0.70	0.91	0.79
fastText-BOT	0.87	0.70	0.80	0.74
HATT	0.93	0.93	0.95	0.93
Bi-LSTM	0.93	0.86	0.98	0.91
RCNN	0.90	0.86	0.90	0.87
DP-CNN	0.93	0.90	0.90	0.90
CNN-LSTM	0.94	0.93	0.94	0.94
Attentional Bi-LSTM	0.93	0.90	0.98	0.93
A-CNN-LSTM	0.94	0.92	**0.98**	0.94
openAI-Transformer	0.95	0.94	0.96	0.94
DLM	**0.96**	**0.95**	0.97	**0.96**

Table 8: Comparison with baselines in terms of four evaluation metrics

to minimum learning rate divisor, cooldown percentage, maximum momentum and minimum momentum in that order. For the STLR the parameters are maximum to minimum learning rate divisor and cut fract. Cut fract is the fraction of iterations we increase the LR. To obtain a sensible learning rate, the learning rate finder (LRF) introduced by Smith (2017) was used. The hyper-parameters are directly transferred from (Howard and Ruder, 2018).

- **Language Model (LM)** - Batch Size $\rightarrow$ 32, BPTT $\rightarrow$ 70, Gradient Clipping $\rightarrow$ (0.4, 0.12), STLR ratio $\rightarrow$ 32, cut fract $\rightarrow$ 0.1, CLR $\rightarrow$ (10, 10, 0.95, 0.85), Weight Dropout $\rightarrow$ 0.5, LR $\rightarrow$ 0.0001, Weight Decay $\rightarrow$ 0.0000001. The Adam optimizer is used.

- **Medium Model (MM)** - Batch Size $\rightarrow$ 32, BPTT $\rightarrow$ 70, Weight Decay $\rightarrow$ 0.0000001. The model is gradually unfrozen (Howard and Ruder (2018)) by unfreezing the last layer first and then unfreezing all subsequent layers. STLR ratio $\rightarrow$ 32 and a cut fract $\rightarrow$ 0.5 were used after the last layer was unfrozen, and an STLR ratio $\rightarrow$ 20 and a cut fract $\rightarrow$ 0.1 was used when all layers were unfrozen.

- **Disclosure Model (DM)** - Learning Rate $\rightarrow$ 0.3, Batch Size $\rightarrow$ 52, BPTT $\rightarrow$ 70, Weight Decay $\rightarrow$ 0.0000001, Cyclical Learning Rates $\rightarrow$ (10, 10, 0.98, 0.85) are used. The model is gradually unfrozen layer by layer with the same hyper-parameters applied to each layer. The Howard dropouts are applied with a multiplier of 1.8 and no gradient clipping is applied. The Adam optimizer is used.

6 Results and Analysis

6.1 Performance

Table 8 describes the performance of the baseline classifiers as well as the deep learning models based on four evaluation metrics.

The Disclosure Language Model outperforms all baseline models, including RNNs, LSTMs, CNNs, and the linear DAD and SDA models. The A-CNN-LSTM and the Hierarchical Attention Model has a high recall due to its ability to capture long term dependencies better. The attention mechanism allows the model to retain some crucial hidden information when the sentences are quite long. GRUs perform poorly as they are unable to learn some latent features of the sequence that are not directly tied to the elements of the sequence. VD-CNN models typically require a large

dataset for effective feature extraction. Nine layers were used, as going deeper decreased the accuracy. Short-cut connections tend to help reduce degradation. CL-CNNs may generate unusual words as they would suffer from a higher perplexity due to the nature of prediction (character-by-character). Also, longer training time can lead to vanishing gradients. The fastText model can generate embeddings quicker but performs similarly to the Char-CNN model.

The AWD-LSTM architecture used in the Disclosure Language Model can avoid catastrophic forgetting. The main benefit, however of the ULMFiT based Disclosure Language Model is that it can perform classifier re-training with a minimal amount of data. The openAI-Transformer model comes a close second in terms of performance. The results show that augmenting the training data with additional domain-specific data (i.e., Tweets) helps to obtain better F1-scores for the segregation of tweets containing instances of personal experiences of sexual harassment.

6.2 Error Analysis

An analysis has been done to show which texts lead to erroneous and a possible explanation of why that might have been the case.

- **Non-Serious -** *"I got raped at FIFA the last time I played lol"* has a flippant tone. However, the model predicted this as Disclosure because of lack of more contextual information.

- **Third person quote -** *"I was followed and harassed by two guys on my way back home last night." This is what my friend had to say after spending one day in Baja.* Here someone is referring to another person's recollection. However, this text contains all the linguistic markers associated with assault disclosure.

- **Uncertainty-** 1.*"He was my teacher and I was 12. #metoo"*.

  2. *"I too am a metoo survivor"*

  Here, the system cannot pick up the context as there is no explicit mention of assault or harassment.

- **Unfamiliarity-** *"I was walking home, and I saw in broad daylight a man walking towards me furiously rubbing his privates looking at me"*. The current dataset lacks in terms of a broad range of phrases that can imply sexual harassment.

- **First person account-** *"senatorcollins i beg you for my 12 year old daughter who was sexually assaulted by her teacher please do not vote yes on kavanaugh"*. The sentence, although in the first person, refers to someone else's experience.

- **Tweets based on a specific current event-** *"I believe Dr. Ford because the same thing happened to me"*. The user assumes that a majority of the readers will be able to gather context from the amount of information provided. However, the system is unable to pick up this nuance because of lack of information about current events.

7 Ethical Considerations

Human language processing, and human language touches many parts of life, these areas also have an ethical dimension. For example, languages define linguistic communities, so inclusion and bias become relevant topics. Based on the issues highlighted in (Schmaltz (2018)), we address these as:

- **Privacy:** Individual consent from users was not sought as the data was publicly available and attempts to contact the author for research participation could be deemed coercive and may change user behavior.

- **Fairness, Bias & Discrimination:** The exhaustive nature of training data introduces bias in terms of how representative the dataset and hence the trained model is of an underlying community. While it's not possible to capture all demographics, we try to maximize our coverage by building our dataset in two phases by first developing a lexicon from various microblogging sites.

- **Interpretation:** Although our work attempts to analyze aspects of users' nuanced and complex experiences, we acknowledge the limitations and potential misrepresentations that can occur when researchers analyze social media data, particularly data from a vulnerable population or group to which the researchers do not explicitly belong. The main

aim of this study was to determine whether it was possible to categorize tweets in this way, rather than to assume the coding was accurate immediately.

8 Conclusion and Future Work

In this work, we proposed a Disclosure Language Model, a three-part ULMFiT architecture, for the task of analyzing disclosures of sexual harassment on social media. On a manually annotated real-world dataset, created in two steps to capture a broad demographic, our systems could often achieve significant performance improvements over (i) systems that rely on handcrafted textual features and (ii) Generic deep learning based systems. An extensive comparison shows the merit of using Medium-Specific Language Models based on an AWD-LSTM architecture, along with an augmented vocabulary which is capable of representing deep linguistic subtleties in the text that pose challenges to the complex task of sexual harassment disclosure. Our future agenda includes: (i) developing a medium-agnostic model robust to the changes in linguistic styles over various forms of social media, (ii) exploring the applicability of our analysis and system to identifying patterns and potential prevention and (iii) applying social network analysis to leverage community interaction and get an overall better understanding.

References

Nazanin Andalibi, Oliver L Haimson, Munmun De Choudhury, and Andrea Forte. 2016. Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 3906–3918. ACM.

Farzindar Atefeh and Wael Khreich. 2015. A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1):132–164.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee.

Katherine Bogen, Kaitlyn Bleiweiss, and Lindsay M. Orchowski. 2018. Sexual violence is notokay: Social reactions to disclosures of sexual victimization on twitter. *Psychology of Violence*.

Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2016. Very deep convolutional networks for text classification. *arXiv preprint arXiv:1606.01781*.

Jeremy Howard and Sebastian Ruder. 2018. Fine-tuned language models for text classification. *arXiv preprint arXiv:1801.06146*.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. *arXiv e-prints*, page arXiv:1801.06146.

Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 562–570.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

Sweta Karlekar and Mohit Bansal. Unc chapel hill 1swetakar, mbansall@ cs. unc. edu.

Aparup Khatua, Erik Cambria, and Apalak Khatua. 2018. Sounds of silence breakers: Exploring sexual violence on twitter. pages 397–400.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.

Bun-Hee Lee. 2018. # me too movement; it is time that we all act and participate in transformation. *Psychiatry Investigation*, 15(5):433–433.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*.

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 142–150. Association for Computational Linguistics.

Debanjan Mahata, Jasper Friedrichs, Rajiv Ratn Shah, and Jing Jiang. 2018. Detecting personal intake of medicine from twitter. *IEEE Intelligent Systems*, 33(4):87–95.

Debanjan Mahata, John R Talburt, and Vivek Kumar Singh. 2015. From chirps to whistles: discovering event-specific informative content from twitter. In *Proceedings of the ACM web science conference*, page 17. ACM.

Lydia Manikonda, Ghazaleh Beigi, Subbarao Kambhampati, and Huan Liu. 2018. *metoo Through the Lens of Social Media*, pages 104–110.

Lydia Manikonda, Venkata Vamsikrishna Meduri, and Subbarao Kambhampati. 2016. Tweeting the Mind and Instagramming the Heart: Exploring Differentiated Content Sharing on Social Media. *arXiv e-prints*, page arXiv:1603.02718.

Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018. Detecting offensive tweets in hindi-english code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26.

Natalie McClain and Angela Frederick Amar. 2013. Female survivors of child sexual abuse: Finding voice through research participation. *Issues in Mental Health Nursing*, 34(7):482–487.

Mayank Meghawat, Satyendra Yadav, Debanjan Mahata, Yifang Yin, Rajiv Ratn Shah, and Roger Zimmermann. 2018. A multimodal approach to predict social media popularity. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 190–195. IEEE.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and optimizing lstm language models. *arXiv preprint arXiv:1708.02182*.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and Optimizing LSTM Language Models. *arXiv e-prints*, page arXiv:1708.02182.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pages 1310–1318.

Rajib Rana. 2016. Gated recurrent unit (gru) for emotion classification from noisy speech. *arXiv preprint arXiv:1612.07778*.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180. Association for Computational Linguistics.

Kristian Rother and Achim Rettberg. 2018. Ulmfit at germeval-2018: A deep neural language model for the classification of hate speech in german tweets.

Ramit Sawhney, Prachi Manchanda, Puneet Mathur, Rajiv Shah, and Raj Singh. 2018a. Exploring and learning suicidal ideation connotations on social media with deep learning. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 167–175.

Ramit Sawhney, Prachi Manchanda, Raj Singh, and Swati Aggarwal. 2018b. A computational approach to feature extraction for identification of suicidal ideation in tweets. In *Proceedings of ACL 2018, Student Research Workshop*, pages 91–98.

Allen Schmaltz. 2018. On the utility of lay summaries and ai safety disclosures: Toward robust, open research oversight. In *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*, pages 1–6.

Nicolas Schrading, Cecilia Alm, Ray Ptucha, and Christopher Homan. 2015. An analysis of domestic abuse discourse on reddit. pages 2577–2583.

Rajiv Shah and Roger Zimmermann. 2017. *Multimodal analysis of user-generated multimedia content*. Springer.

Leslie N Smith. 2017. Cyclical learning rates for training neural networks. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 464–472. IEEE.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Duc-Thuan Vo, Vo Thuan Hai, and Cheol-Young Ock. 2015. Exploiting language models to classify events from twitter. *Computational intelligence and neuroscience*, 2015:4.

Jenq-Haur Wang, Ting-Wei Liu, Xiong Luo, and Long Wang. 2018. An lstm approach to short text sentiment classification with word embeddings. In *Proceedings of the 30th Conference on Computational Linguistics and Speech Processing (ROCLING 2018)*, pages 214–223.

Christine Wekerle, Negar Vakili, Sherry Stewart, and Tara Black. 2018. The utility of twitter as a tool for increasing reach of research on sexual violence. *Child abuse neglect*, 85.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1480–1489.

Hang Yuan, Jin Wang, and Xuejie Zhang. 2018. Ynuhpcc at semeval-2018 task 11: Using an attention-based cnn-lstm for machine comprehension using commonsense knowledge. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1058–1062.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.

Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis Lau. 2015. A c-lstm neural network for text classification. *arXiv preprint arXiv:1511.08630*.