#YouToo? Detection of Personal Recollections of Sexual Harassment on Social Media

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

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

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

Debanjan Mahata
Bloomberg
dmahata@bloomberg.net

Abstract

The availability of large-scale online social data, coupled with computational methods, can help us answer fundamental questions relating to our social lives, particularly our health and well-being. The #MeToo trend has led to people talking about personal experiences of harassment more openly. This work attempts to aggregate such experiences of sexual abuse to facilitate a better understanding of social media constructs and to bring about social change. It has been found that disclosure of abuse has positive psychological impacts. Hence, we contend that such information can be leveraged to create better campaigns for social change by analyzing how users react to these stories and can be used to obtain a better insight into the consequences of sexual abuse. We use a three-part Twitter-Specific Social Media Language Model to segregate personal recollections of sexual harassment from Twitter posts. An extensive comparison with state-of-the-art generic and specific models along with a detailed error analysis explores the merit of our proposed model.

1 Introduction

Global estimates indicate that about 1 in 3 women worldwide has experienced either physical and/or sexual intimate partner violence or non-partner sexual violence in their lifetime [1]. The hashtag #MeToo has been prevalent on Twitter as a campaign centered around sharing stories of sexual harassment in the act of solidarity with other victims and spreading awareness of a widespread and endemic issue. With vast amounts of people sharing their recollections of sexual harassment on the Internet, it is important that we make scientific use of this data to increase awareness and enable real-world change. Manually sorting and comprehending the information shared in these stories is an arduous task. Hence our work can serve as the missing link between online activism and real change.

Health information seeking and sharing practices online have been known in helping people cope with mental health problems (De Choudhury and De, 2014). Studies have shown that online forums and support groups provide a conducive environment allowing people to get connected with others who share similar stories, thus act as a path to obtaining help and advice around mental health problems (Eysenbach et al., 2004). Moreover, self-disclosure is therapeutic for mental health communities (Johnson and Ambrose, 2006).

Our study proposes a Twitter-Specific Social Media Language Model for the aggregation of tweets containing personal stories of sexual harassment. Manikonda et al. (2018) have carried out a preliminary analysis of the user engagement, discussion topics, word connotations, and sentiment concerning the #metoo movement. Andalibi et al. (2016) have explored anonymity and support seeking during the #metoo movement. However, very few studies have attempted to separate texts containing discussions about sexual harassment from texts containing personal stories of sexual harassment experiences. Efforts have been made to aggregate domestic abuse stories from Reddit by Schrading et al. (2015). Karlekar and Bansal (2018a) have attempted to categorize personal stories into categories like *ogling, commenting, groping*. Our study aims to help this body of research grow by automating the process of collection of tweets containing recollections of sexual harassment.

* denotes equal contribution.

[1] https://www.who.int/news-room/fact-sheets/detail/violence-against-women

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2527–2537
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

1.1 Clinical Perspective

Prior research in psychology has demonstrated the importance of social support in combating depression (George et al., 1989). It is argued that social intimacy, social integration, nature of social networks as well as the individual perception of being supported by others are important and essential to quick recovery from mental health problems (Caplan and Turner, 2007). The internet is increasingly used for seeking and sharing health information online, and such activity is known to have connections to health-related behaviors ((Sillence et al., 2007); (Liu et al., 2013)). Online support groups are popular sources of information and support for many internet users (White and Dorman, 2001). These forums tend to be very different from similar offline groups; for instance, people are likely to discuss problems that they do not feel comfortable discussing in person (Johnson and Ambrose, 2006). Moreover, such online health communities are known to foster well-being, a sense of control, self confidence, social interactions, and improved feelings.

In the context of mental health in particular, Moreno et al. (2011) demonstrated that status updates on Facebook could reveal symptoms of major depressive episodes, while Park et al. (2013) found differences in the perception of Twitter use between depressed and non-depressed users: the former found value in Twitter due to the ability to garner social awareness and engage in emotional interaction.

We have presented, to the best of our knowledge, the first comprehensive dataset and methodology for detection of personal stories of sexual harassment on Twitter. We carry out extensive comparisons of our proposed Medium Specific Social Media Language Model with respect to baselines. Our work may provide a wealth of resources to clinicians, health practitioners, caregivers, and policy makers to identify communities at risk.

2 Related Work

Natural language processing (NLP) techniques can be used to make inferences about peoples mental states from what they write on Facebook, Twitter, and other social media. These inferences can then be used to create online pathways to direct people to health information and assistance and also to generate personalized interventions. Regrettably, the computational methods used to collect, process, and utilize online writing data, as well as the evaluations of these techniques, are still dispersed in the literature. 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. Sawhney et al. (2018) have worked on the detection of posts containing suicidal ideation on Twitter. 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 specifically 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 (2018b) 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.

A growing body of work has demonstrated that social media is an increasingly adopted platform allowing users to communicate around a variety of health concerns ((Paul and Dredze, 2011); (Andalibi et al., 2016)). Newman et al. (2011) interviewed people with significant health concerns who participated in both OHCs and Facebook. Oh et al. (2013) examined peoples use of Facebook for health purposes and showed that emotional support was a significant predictor of health self-efficacy. 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 addressed to capture subtle nuances and understand the context better.

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*
#sexualharassment	*#notallmen*
#mentoo	*#timesup*
#womensreality	*#EverydaySexism*

Table 1: Words/Phrases linked with Sexual Harassment

3 The Sexual Harassment Recollection (SHR) Dataset

3.1 Data Collection

One of the foremost challenges with detecting personal recollections of sexual harassment is the lack of availability of a public dataset due to privacy and anonymity concerns borne out of social stigma associated with sexual harassment. Motivated by the need to create a fresh dataset, a corpus of words and phrases were developed using anonymized data from known Sexual Harassment forums. Between November 2016 and December 2018, these forums were scraped for the user posts and human annotators were asked to identify if these posts were related to sexual harassment. In addition to this, user posts (containing tags of sexual abuse) from the micro-blogging websites Reddit were collected and added to this collection. These were subsequently human annotated based on them containing personal recollections of sexual harassment or not. Then, Term Frequency/Inverse Document Frequency (TF-IDF) method was applied to this set of manually annotated texts to identify terms which frequently appear in the texts belonging to the Recollection class and less frequently in the Non-Recollection class. These terms play a role in differentiating between the two classes. Finally, manual annotators were asked to remove any terms from this list, which were not based on sexual harassment as well as duplicate terms. This gave a final lexicon of 70 terms consisting of but not limited to the phrases/words of Table 1. The public Streaming API offered by Twitter allows programmatic collection of tweets as they occur, filtered by specific criteria. Using the same, anonymized data was collected from Twitter.

The tweets retrieved from Twitter using the API contain extraneous information. It can be associated with a URL, user mention, media files(image, audio, and video), timestamp, number of retweets. For the tasks in this paper, the text from each tweet was extracted while the rest of the information about the tweet was discarded. Although the tweets were collected from the 'stream' based on sexual harassment earlier developed, the exact sentiment of the tweets was unknown. Tweets about sexual harassment could be related to other things as well. *Eg. Sexual harassment awareness campaign and prevention, a news report, sarcasm etc.* This made a manual annotation of the dataset imperative for better accuracy.

3.2 Data Annotation

The final dataset consisting of 5119 text sentences from different tweets was then, manually annotated by two humans, one, a student of gender studies, the other, a student of clinical psychology and an outside annotator (a teacher of gender studies and a non activist feminist) who helped the annotators with conflicts, reviewing our annotations to mitigate bias and confusions. Tweets annotated as *Recollection* are labeled as 1 and the rest of the tweets are labeled 0.

To reliably identify disclosure, we clearly define a tweet to be labeled as Recollection if it explicitly mentions a personal recollection of sexual harassment; e.g. *I was molested by ex-boyfriend. Sexual Harassment* in our case entails a broader definition of this term, which includes sexual abuse, sexual assault, rape and sexual harassment. The remaining tweets not marked as Recollection belonged to one of the following categories:

- Awareness related tweets; e.g. *Do you know what the consequences of domestic violence include? Learn more here ¡url¿ #feminism #meToo*
- Flippant references; e.g. *Dude, I can't play Fortnite! I got raped there, haha meToo*
- News reports and incidents; e.g. *In an exclusive interview with BBC Asian Network, bollywood superstar @iamsrk speaks about the #meToo movement, film censorship and #Brexit.*
- Tweets describing other's experiences; e.g. *My best friend was sexually assaulted. #meToo #assault*
- Tweets using #meToo in a different context; e.g. *Yumm! I'm starving for spring rolls too #meToo #chinese*
- Other remaining tweets; e.g., *So the #meToo movement doesn't apply to democrats? Oh ok, got it.*, *Exploiting the #meToo movement for political gain? Not cool.*

Finally, after an agreement between the annotators (using majority decision to label the mixed cases), 1126 tweets in the dataset (22% of the dataset) were annotated as *Recollection* with an average value of Cohen Kappas inter-annotator agreement $\kappa = 0.83$, while the rest fell into the category of Discussion. Our dataset will be made publicly available, following the guidelines mentioned in Section 7 to facilitate further research and analysis on this very pertinent issue [2].

3.3 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.
- Stopwords corpus was taken from NLTK and was used to eliminate words which provide little to no information about individual tweets (Loper and Bird, 2002).
- URLs, screen names(username), hashtags(#), digits(0-9), and all Non-English words were removed from the dataset [3]

4 The Social Media Language Model (SMLM)

Our work considers deep learning techniques for the detection of social media disclosures of sexual harassment. The majority of methods used to study NLP problems employing shallow machine learning models and time-consuming, hand-crafted features suffer from dimensionality problems since linguistic information is usually represented with sparse representations (high-dimensional features). (Khatua et al., 2018). Bag-of-words approaches tend to have high recall but lead to high rates of false positives because lexical detection methods classify all messages containing particular terms only.

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.

A language model that possesses universal properties could be useful in cases where there is a lack of annotated datasets or language resources, which is prevalent in NLP research. We propose a three-part 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

[2] github.com/arijit1410/ACL2019-YouToo

[3] https://abiword.github.io/enchant/

Text	Label
# WhenIWas 15 I was molested by my best friend	1
I was sexually assaulted by my step brother in 2009.	1
At 8 years old, an aldult family member sexually assaulted me.	1
I was 7 the first time I was sexually assaulted.	1
I was sexually assaulted by at least 3 different babysitters by the time I was 6 years old.	1
#Me too campaign stop sexual harassment and sexual assault.	0
Trying to silence sexual assault victims is another one. The list goes on and on	0
Then call for people that cover up sexual assault like Jim Jordan to resign???	0
sexual assault on public transport is real	0
agreed! metoo is not just exclusively for women!	0

Table 2: Example tweets from the annotated dataset

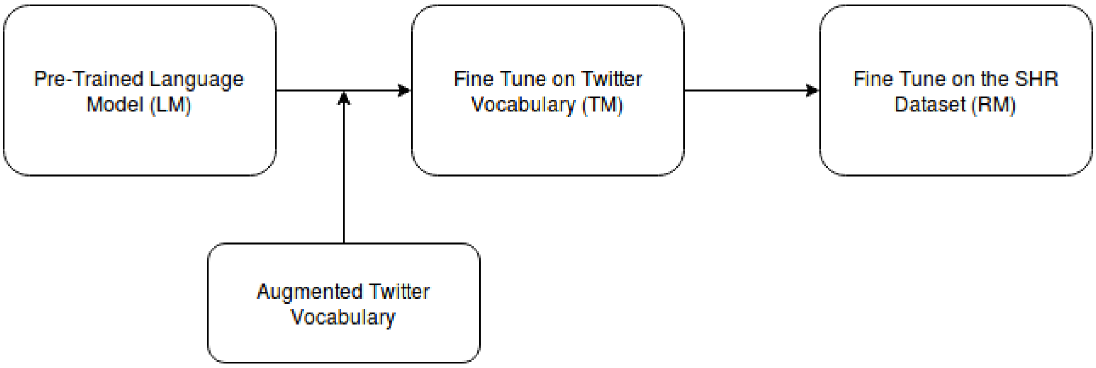
Pre-Trained Language Model (LM)
Fine Tune on Twitter Vocabulary (TM)
Fine Tune on the SHR Dataset (RM)
Augmented Twitter Vocabulary

Figure 1: The Social Media Language Model Overview

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.

4.1 Classification

The Language Model (LM) is trained from a large corpus of unlabeled data. In this case a pretrained Wikipedia Language Model was used. This Language Model is then used as the basis to train a Twitter Model (TM) 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 retrained on a subset of the Twitter Sentiment140 dataset [4]. This augmented vocabulary improves the model's domain understanding of Tweet syntax and semantics. Finally, a binary classifier is trained on top of the Twitter Model from a labeled dataset. This approach facilitates the reuse of pre-trained models for the lower layers.

5 Experiment Setup

5.1 Baselines

In order to make a fair comparison between all the models mentioned above, the experiments are conducted with respect to certain 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 including: RNN (Liu et al., 2016), LSTM/Bi-LSTM (Merity et al., 2017), CNN (Kim, 2014), Character-Level Convolutional Network (CL-CNN) (Zhang et al., 2015), fastText (Joulin et al., 2017), Hierarchical Attention Networks (HATT) (Yang et al., 2016), and an Atten-

[4] https://www.kaggle.com/kazanova/sentiment140

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
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
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
SMLM	**0.96**	**0.95**	0.97	**0.96**

Table 3: Performance Comparisons on the SHR Dataset

Task (Twitter)	Architecture	Accuracy	Precision	Recall	F1
Ours	SMLM + No Augmented Vocab	0.92	0.86	0.95	0.90
Ours	SMLM + Augmented Vocab	0.96	0.95	0.97	0.96
Stance Detection	SMLM + No Augmented Vocab	0.52	0.52	0.54	0.51
Stance Detection	SMLM + Augmented Vocab	0.67	0.66	0.63	0.64
Hate Speech	SMLM + No Augmented Vocab	0.90	0.90	0.91	0.90
Hate Speech	SMLM + Augmented Vocab	0.93	0.93	0.94	0.93

Table 4: Variation in performance with the inclusion of augmented vocabulary on Twitter Datasets

tion Based CNN-LSTM (A-CNN-LSTM) (Yuan et al., 2018). The Transformer based Language Model (Vaswani et al. (2017) ; Ritter et al. (2010)) was used to compare the performance of Language Model based architectures.

For RNN and LSTM, pre-trained Glove word embeddings which were trained on 2 billion tweets are used as features for classification. ReLU activation function (Nair and Hinton, 2010) was used for the CNN layers in the CNN-LSTM Models. A dropout probability of 0.2 was used. The batch size was chosen to be 64, and a total number of epochs were 25. The Adam optimizer was used for all the models (Kingma and Ba, 2014) along with a learning rate of 0.001. A small subset (10%) of the dataset is held back for testing on unseen data.

5.2 SMLM Architectures and Parameters

Our method uses the Weight Dropped AWD-LSTM architecture (Merity et al., 2017). 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 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 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. The dropout used by Howard and Ruder (2018) are (Input Layer →

0.25, General Layer → 0.1, LSTM Internal → 0.2, Embedding Layer → 0.02, Between LSTM Layers → 0.15).

- **Language Model (LM)** - Batch Size → 32, BPTT → 70, Gradient Clipping → (0.4, 0.12), STLR ratio → 32 ,cut fract → 0.1, CLR → (10, 10, 0.95, 0.85). Weight Dropout → 0.5. The Adam optimizer has been used.

- **Twitter Model (TM)** - Batch Size → 32, BPTT → 70, Weight Decay → 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 → 32 and a cut fract → 0.5 were used after the last layer was unfrozen, and an STLR ratio → 20 and a cut fract → 0.1 was used when all layers were unfrozen.

- **Recollection Model (RM)** - Learning Rate → 0.3, Batch Size → 52, BPTT → 70, Weight Decay → 0.0000001, Cyclical Learning Rates → (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.

5.3 Further Exploration

The Twitter Model(TM) in our proposed method enables fine-tuning of the language model on a large corpus of domain-specific data. To validate that the model is generic, and to show that the addition of an augmented vocabulary boosts the performance of classifiers across other tasks as well, where the training dataset is relatively small, we compare the performance of the SMLM on several publicly available small datasets, with and without the extended vocabulary. The Political Stance Detection Dataset (SemEval-2016 Task 6) uses a small dataset of 4163 tweets for classification [5]. The labeled data provided consists of a target topic, a Tweet that pertains to it, and stance of the Tweet towards the target. The data is already split into a training set (containing 2,914 Tweets) and a test set (containing 1,249 Tweets). We also test the SMLM model on the Twitter Hate-Speech dataset created by Davidson et al. (2017) [6]. The tweets are labelled as *hate-speech, offensive* and *neutral*. We augment the language model with the same subset of 100,000 tweets that we have used for the SMLM model.

6 Results and Analysis

6.1 Performance

Table 3 describes the performance of the baseline classifiers as well as the deep learning models based on four evaluation metrics.

The Social Media 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 better capture long term dependencies. The attention mechanism allows the model to retain some important hidden information when the sentences are quite long. 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 is able to generate embeddings quicker but performs similarly to the CL-CNN model. The AWD-LSTM architecture used in the Social Media Language Model is able to avoid catastrophic forgetting. The main benefit, however of the ULMFit based Social Media Language Model is that it can perform classifier re-training with a very limited amount of data. The openAI-Transformer model comes a close second in terms of performance.

6.2 Generic Nature of the SMLM Model

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. Table 4 shows that addition of this augmented vocabulary can be extended to other tasks on twitter also, with limited training data, implying that our proposed model has the potential to be generic across other medium specific tasks as well.

We make the following observations.

- Our fine-tuned language model can generalize to the unstructured and messy language syntax of Tweets.

[5] http://alt.qcri.org/semeval2016/task6/

[6] https://data.world/crowdflower/hate-speech-identification

- The SMLM model can achieve an improved F1 score with minimal task-specific customization for each model and with limited computing resources.

6.3 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 (Table 5). L is the correct label, and M is label predicted by the SMLM model.

- **T1**: This text has a flippant tone. However, the system cannot pickup this nuance because it does not understand the casual nature of the discussion and the misplaced use of the term ”rape”.

- **T2**: Here, someone is referring to another person’s recollection. However, this text contains all the linguistic markers associated with assault disclosure.

- **T3**: Here, readers can pick up the context of this being a probable recollection of sexual harassment by a teacher when the author was 12. The system cannot pickup the context, the same way a human can, based on previous trends in other tweets.

- **T4**: The system cannot pickup the meaning of the word ”metoo survivor”. A human can associate the term ”survivor” with ”metoo” if they have context from other tweets in which people talk about they have survived sexual abuse and harassment.

- **T5** The current training dataset lacks in terms of a broad range of phrases that can imply sexual harassment.

- **T6**: The sentence, although in the first person, refers to someone else’s experience.

- **T7**- In this case, 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. Specifically, the system does not have prior information on who Dr. Ford is.

7 Ethical considerations and limitations

Research with sexual assault victim-survivors can present heightened ethical challenges. This means that research on this topic must be handled with particular skill, care and respect. We address the following limitations :

- **Confidentiality:** 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.

  For instance, some victims may be deterred from coming forward if they knew they are being ”tracked” by algorithms.

- **Justice:** 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.

  Any potential benefits of a project should be balanced carefully against the potential to cause harm. If bias is present, the benefits of the research are not shared across the community.

- **Potential Misrepresentation:** 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. Further note that by no means the goal of this research is to claim that our coding is accurate, we only attempt to study whether it is possible to categorize tweets in this way.

  Particular care was taken to ensure all members of the research team have been extensively trained in undertaking research sensitively, and are aware of relevant ethical issues.

8 Conclusion and Future Work

In this work, we proposed a Social Media Language Model, a three part ULMFiT architecture,

Id	Tweet	L	M
T1	*"Dude, I can't play Fortnite! I got raped there, haha meToo"*	0	1
T2	*"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.*	0	1
T3	*"He was my teacher and I was 12. #metoo"*	1	0
T4	*"I too am a metoo survivor"*	1	0
T5	*"I was walking home and I saw in broad daylight a man walking towards me furiously rubbing his privates looking at me".*	1	0
T6	*"senatorcollins i beg you for my 12 year old daughter who was sexually assaulted by her teacher please do not vote yes on kavanaugh".*	0	1
T7	*"I believe Dr Ford because the same thing happened to me"*	1	0

Table 5: Error Analysis

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 large demographic, our systems could often achieve significant performance improvements over systems that rely on handcrafted and textual features and 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 text that pose challenges to the complex task of detecting sexual harassment disclosure. We also hope this study enables further research in terms of how people seek support online on sexual harassment and mental health-related problems. Our future agenda includes exploring the applicability of our analysis and system for identifying patterns and potential prevention. We also plan to use this model to solve other downstream medium-specific tasks pertaining to mental health and welfare.

References

Nazanin Andalibi, Oliver Haimson, Munmun De Choudhury, and Andrea Forte. 2016. Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. pages 3906–3918.

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*.

Scott E Caplan and Jacob S Turner. 2007. Bringing theory to research on computer-mediated comforting communication. *Computers in human behavior*, 23(2):985–998.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Gunther Eysenbach, John Powell, Marina Englesakis, Carlos Rizo, and Anita Stern. 2004. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. *Bmj*, 328(7449):1166.

Linda K George, Dan G Blazer, Dana C Hughes, and Nancy Fowler. 1989. Social support and the outcome of major depression. *The British Journal of Psychiatry*, 154(4):478–485.

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.

Grace J Johnson and Paul J Ambrose. 2006. Neo-tribes: The power and potential of online communities in health care. *Communications of the ACM*, 49(1):107–113.

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. 2018a. Safecity: Understanding diverse forms of sexual harassment personal stories. *arXiv preprint arXiv:1809.04739*.

Sweta Karlekar and Mohit Bansal. 2018b. 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*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Leslie S Liu, Jina Huh, Tina Neogi, Kori Inkpen, and Wanda Pratt. 2013. Health vlogger-viewer interaction in chronic illness management. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 49–58. ACM.

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

Edward Loper and Steven Bird. 2002. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*.

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.

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

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.

Megan A Moreno, Lauren A Jelenchick, Katie G Egan, Elizabeth Cox, Henry Young, Kerry E Gannon, and Tara Becker. 2011. Feeling bad on facebook: depression disclosures by college students on a social networking site. *Depression and anxiety*, 28(6):447–455.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, USA. Omnipress.

Mark W Newman, Debra Lauterbach, Sean A Munson, Paul Resnick, and Margaret E Morris. 2011. It's not that i don't have problems, i'm just not putting them on facebook: challenges and opportunities in using online social networks for health. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 341–350. ACM.

Hyun Jung Oh, Carolyn Lauckner, Jan Boehmer, Ryan Fewins-Bliss, and Kang Li. 2013. Facebooking for health: An examination into the solicitation and effects of health-related social support on social networking sites. *Computers in human behavior*, 29(5):2072–2080.

Minsu Park, David W McDonald, and Meeyoung Cha. 2013. Perception differences between the depressed and non-depressed users in twitter. In *Seventh International AAAI Conference on Weblogs and Social Media*.

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.

Michael J Paul and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. In *Fifth International AAAI Conference on Weblogs and Social Media*.

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.

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

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

Elizabeth Sillence, Pam Briggs, Peter Richard Harris, and Lesley Fishwick. 2007. How do patients evaluate and make use of online health information? *Social science & medicine*, 64(9):1853–1862.

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.

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.

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.

Marsha White and Steve M Dorman. 2001. Receiving social support online: implications for health education. *Health education research*, 16(6):693–707.

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. Ynu-hpcc 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.