Sexual Harassment Detection

Shreya Reddy[1], Pedram Khayyatkhoshnevis[1], Abhigya Koirala[1], and Shrinal Thakkar[1]

[1]Department of Computer Science, Lakehead University

***Abstract*—Social media has become an inevitable part of our everyday lives. Sexual harassment is common on social media, particularly directed to female users. In October 2017, the #MeToo movement became viral on Twitter that allowed millions of female users to share their stories online. In this paper, we attempt to replicate a state-of-the-art model and extend that work by developing a CNN, LSTM,CNN-LSTM, and BiLSTM neural models to detect abusive stories on Reddit and Twitter. We used sentiment classification combined with feature extraction method, GloVe. We also implemented all the necessary steps such as a web crawler tool, pre-processing and post-processing methods. Based on evaluation data our model achieved 97% accuracy and 97% F-1 score.**

I. Introduction

With the emergence of smartphones, people are able to engage with social media services such as Facebook[1], Twitter[2], and WhatsApp[3] that allows them to share their ideas and opinions [1]. This usage of social media has created an impact on people's day to day life. According to research, as of January 2020 [2], nearly 60% of the world population is using some kind of online services. It also states that nearly 49% of internet users are active on social media on every given day [3]. With an enormous number of active users, a large amount of data is generated. People nowadays post regarding different topics on social media which range from their views and opinions on different domains such as politics, health, and lifestyle to discuss their personal life. Such posts can be very useful for analyzing customer behavior, certain social aspects, and other domain-specific studies that are a subject of interest by corporations, policy-makers, and government agencies [1]. Furthermore, it can help to solve social problems such as sexual harassment. Sexual harassment, with a long history, is a widespread problem worldwide. Statistics show that women and girls are held at high risk of abuse. One in three women experiences physical, verbal or sexual harassment [3]. Social media can be a more open and transparent platform for those who have endured abuse to be encouraged to express their traumatic experiences freely and create awareness among the people. The trending #MeToo movement which was started on Twitter illustrates how sharing sexual harassment cases on social media platforms [4] can increase awareness and empower women. This has been one of the major issues around the world. There has been a little work done in detecting sexual harassment on social media. In this study, we focused on the sexual harassment stories from Reddit[4] and Twitter. By using Natural Language Processing (NLP) and supervised classifiers, we classify the stories into 2 groups: harassment and non-harassment. And then we classify the different types of online harassment stories into 2 categories: indirect harassment and sexual harassment. The machine learning algorithms which we used for the above-mentioned tasks include Convolutional Neural Network (CNN), Long short-term memory (LSTM), a combination of both, and Bidirectional LSTM (BiLSTM). The choice of a robust classifier to detect this kind of content justifies testing of all of those different methods.

II. Related Work

This section focuses on a few related topics such as sexual harassment detection, uncertainty detection, irony detection and tokenization methods to extract meaningful and useful language units.

A. Sexual harassment detection

A recent study focused on the hashtag #MeToo movement. The #MeToo movement has been trending on Twitter for quite some time. Researchers in this paper, have proposed a Twitter-specific Social Media Language Model (SMLM) to aggregate the tweets related to personal sexual harassment stories. Some researchers have carried out a preliminary analysis of the user engagement, word connotations and sentiment concerning the #MeToo movement [1]. The researchers in this paper found out that there were few types of research and work done on separating the texts containing discussion about sexual harassment from the texts containing the story of personal sexual harassment experiences. Another attempt has focused on categorizing personal stories into different categories like ogling, commenting, groping [3]. This paper comprises of two important parts: the creation of The Sexual Harassment Recollection (SHR) Dataset and the experimental setup of The Social Media Language Model (SMLM). The SMLM uses deep learning techniques for the detection of sexual harassment in social media. They used the three-part classification method based on Universal Language Model Fine-tuning (ULMFiT) architecture. A pertained Wikipedia Language Model was then used to train the twitter model. A binary classifier was then trained on top of the Twitter Model from a labeled dataset. The proposed method was then compared with other methods such as RNNs, LSTMs, CNNs. The SMLM outperformed all these models in their baseline. It was found that the training data, when augmented with

[1]https://www.facebook.com/
[2]https://twitter.com/
[3]https://www.whatsapp.com/
[4]https://www.reddit.com/

additional domain-specific data, helps to obtain a better F1 score.

B. Uncertainty Detection and Irony detection

Another important aspect of text analysis is differentiating between uncertain and factual statements. This process helps to separate factual sounding tweets from uncertain tweets. Uncertainty cues used in literature are domain and genre dependent. Uncertainty cue detection and disambiguation are one of the main contributions of this research. In their research, uncertainty cue detection is approached as a token labeling problem. Tokens are labeled in three ways, 1) uncertainty cue, 2) inside a token of a cue, or 3) not a part of any cue. As a standard practice, cue-level and sentence level Fβ=1 score was used as an evaluation metric [5]. In order to obtain uniform results, all evaluation datasets were normalized.

Verbal irony generally implies the expression of a positive evaluation when a negative one is intended or vice versa [4]. Figurative language (irony) has significant implications for tasks that treat subjective information, such as sentiment analysis (involves the extraction of positive and negative opinions from online texts). An example of such subjective information is contained in the ironic utterance "Cannot wait to go to the dentist tomorrow!" [4]. We know that the act of going to the dentist is typically an unpleasant situation. This clearly contrasts with the positive expression "cannot wait". Hee et al. take up the challenge of automatically recognizing implicit sentiment in tweets and explores how irony detection benefits from such implicit sentiment information. A lexico-semantic knowledge base and a data-driven approach based on real-time tweets have been used together with an SVM-based irony detection system. An evaluation was conducted against three baselines. Implicit sentiment inferences followed automatically with real-time tweets (accuracy = 72%).

C. Tokenization

While the tokenization method is a well-defined subject, recent studies have improved some aspects of tokenization. A research conducted by Riedl et al. explores the notion of tokenization from low-level character-based token detection to detect the meaning and usefulness of language units [6]. In their research, they aim at identifying units that are composed of many single words that form Multi-Word Expressions (MWE), as well as splitting those units into a single meaningful word. This mechanism uses two ranking methods [6]. The uniqueness rank and the punishment rank. The uniqueness rank is measured by finding the top n similar word to the MWE using the DT dictionary. The punishment ranking finds terms and takes one word (window 1) from the right and left of the term. Then the occurrence of these two words is counted over 1000 samples. If those words often appear with the term in question, then the term is perceived as an incomplete term.

III. Data/ Methodology

In this section, we explain how we gather the training and testing data set and the different methods used for classification.

A. Dataset Collection & Annotation

Reddit is a microblogging website that is a front face of the internet which is a massive collection of forums, where people can share news and content or comment on other people's posts. Reddit contains a substantially large range of forums called subreddits where users post comments that are related to specific topics.

In the recent times, people started sharing their stories about sexual harassment on Reddit, which gives us a good opportunity for extracting those data. We used the Python Reddit API Wrapper (PRAW)[5], a python library that helps scraping data from Reddit. We developed our own dataset since the dataset used in the paper [3] was not made available online. We built a corpus of words and phrases using anonymized data from Reddit sexual harassment forums. We obtained over 200 stories of sexual harassment incidents. The stories include a text description, along with tags of forms of harassment such as sexual abuse, harassment, ogling, commenting, and groping. From this data, we collected some negative data(non-sexual content) which we used it later in our main dataset while collecting our positive data (sexual content) from sentiment140 dataset[6].

The sentiment140 is a dataset that is available on Kaggle[6] and contains 1.6 million tweets that were extracted from the Twitter API. The tweets have been annotated to 0 for negative and 4 for positive and these annotations are used to detect sentiment. The dataset contains 6 fields and this is shown in Table I.

Table I: Sentiment140 dataset fields.

Target	Polarity of the tweet (0 - negative, 2-neutral and 4- positive).
Ids	The id of the tweet.
Date	Date of the tweet that was created.
Flag	The query. If there is no query, then the value is NO_QUERY
User	The user that tweeted.
Text	The text of the tweet.

B. Data Splitting

We used the 70:30 ratio to perform the training and testing split with a random state set to 2003. The sklearn library was used.

C. Data Preprocessing

One of the most important steps in deep learning is data cleaning, rearranging and transforming into understandable data. Since we created our own data, we encountered a lot of noise and irrelevant data that wasn't unnecessary.We used the inbuilt functions provided by the Natural Language Toolkit(NLTK)[7] library for the preprocessing step. The NLTK toolkit contains packages that make machines understand human language and respond appropriately. Table II shows the Reddit dataset before data cleaning.

[5]https://praw.readthedocs.io/

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

[7]https://www.nltk.org/

Table II: Reddit Dataset before cleaning.

Title	Body	Id	....	article
r/SexualHarassm...	If you're interested and willing a...	dzdoq9	....	NaN
My 'fun' experience...	I'm not sure if this would...	a7jyu9	....	NaN
.....	.....	....	....	.....
Star Economist Ro...	NaN	a6cnzf	...	But in interviews with The New...

Below are the types of pre-processing techniques that were used to clean up the data.

1) Normalization: In this step, we converted the given text from uppercase to lowercase, removed punctuation's, white space, strings and special characters.
2) Removal of stopwords: The text contains a lot of stopwords and this is considered as noise in the text. We created a list of stopwords and filtered the list of tokens from these words.
3) Stemming: It's a linguistic normalization that transforms a nonstandard word into its root word. For example, playing or plays becomes "play". We used the PorterStemmer().
4) Lemmatization: A process that reduces words to their base word, generally called lemma. It's quite similar to stemming. But in lemmatization it reduces the inflected word properly. We used the WordNetLemmatizer().
5) URL, score, number of comments, id, subreddit, date created, comments, non-english words were all removed from the Reddit dataset.
   After extracting the Twitter data, we combined the extracted data from Reddit with the positive and neutral tweets from Twitter(sentiment140). The intuition behind this is that the Reddit data was extracted from sexual harassment forums so we can assume that those are sexual harassment-related sentiments. And we assume that the positive and natural tweets from the sentiment140 data set are not about sexual harassment topics. Table III shows the data after combining the positive and negative stories.

Table III: Dataset after cleaning and merging positive and negatives.

Text	**Sentiment**
But in interviews with The New York Times and ...	0
Every year, sexual harassment in the workplace...	0
.....	.....
I have this strange desire to go to confession.....	1
@i_reporter answer sent in dm. try it	1

IV. Experimental Setup

A. Feature Extraction

Word embedding also known as word vectors are numerical representations of words that facilitate language understanding by mathematical operations. They rely on a vector space model that captures the relative similarity of individual word vectors hence providing information on the underlying meaning of words [7]. This is the main reason word embedding are used for text classification and sexual harassment detection. We have used the pre-trained word embedding, Global Vectors for Word Representation (GloVe) which is an unsupervised learning algorithm. This model is imported from the SpaCy library and it contains 6 billion words with each word being a 200 dimensional vector.

B. Models

- **CNN**: For each input, an embedding and a convolutional layer was applied, followed by a max pooling layer. Pretrained word embeddings were used. The convolution features were then passed to a sigmoid layer, which outputs probabilities over two classes.
- **LSTM**: It is a special type of Recurrent Neural Networks (RNN) that is able to capture the long-term dependency among words in short texts [8]. It has three gates: input gate, forget gate and output gate. Our LSTM has 100 hidden units with a batch size of 64 and dropout of 0.2.
- **CNN-LSTM**: We also presented a combined CNN-LSTM architecture. Our CNN-LSTM model consists of an LSTM on top of a CNN. The CNN has 64 filters with kernel size of 5 and embedding dimensions of 200 were used. An LSTM with 200 hidden units was used. The total number of epochs were 7.
- **BiLSTM**: The BiLSTM model comprises two layers of one directional LSTM. The word embeddings for all words in a post are fed to BiLSTM.

Adam optimizer was used for all our models and a learning rate of 0.0001.

V. Results

We evaluated our model using the evaluation metrics, that are F-1 Score, Accuracy, Recall, and Precision. Our model could achieve satisfactory results compared to that of the main paper [3]. The CNN-LSTM model achieved the highest accuracy of 97% accuracy. Table IV shows the classification accuracy of our models: CNN, LSTM, CNN-LSTM, and BiLSTM. The graphical representation of our models are shown in Figure 1, Figure 3 and Figure 5 and other evaluation metrics are shown in Figure 2, Figure 4 and Figure 6.

Table IV: Accuracy Results.

Model	**Accuracy**	**Precision**	**Recall**	**F1**
CNN	0.95	0.94	0.94	0.94
LSTM	0.96	0.96	0.96	0.96
CNN-LSTM	0.97	0.97	0.97	0.97
BiLSTM	0.96	0.97	0.97	0.97

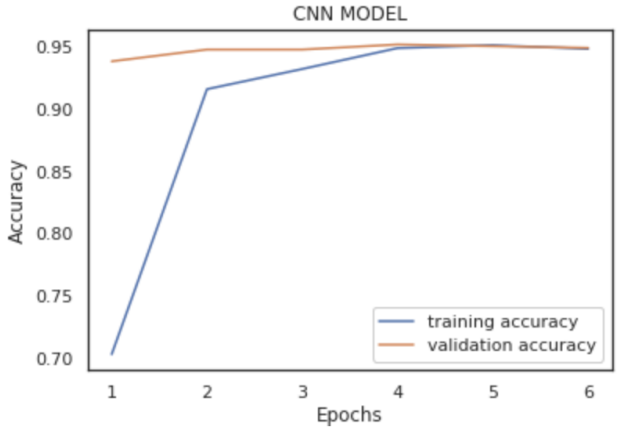

Figure 1: Accuracy for CNN Model.

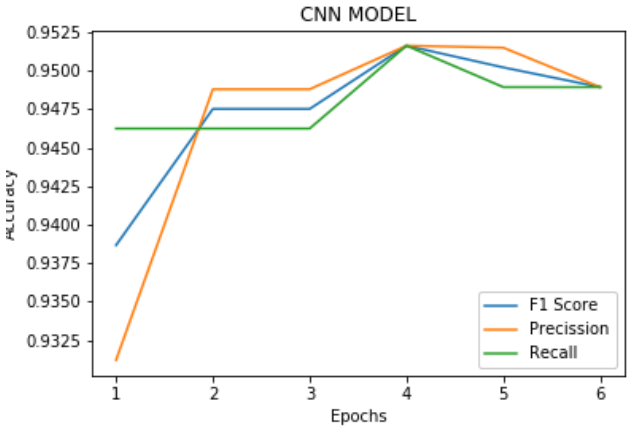

Figure 2: Result for CNN Metrics.

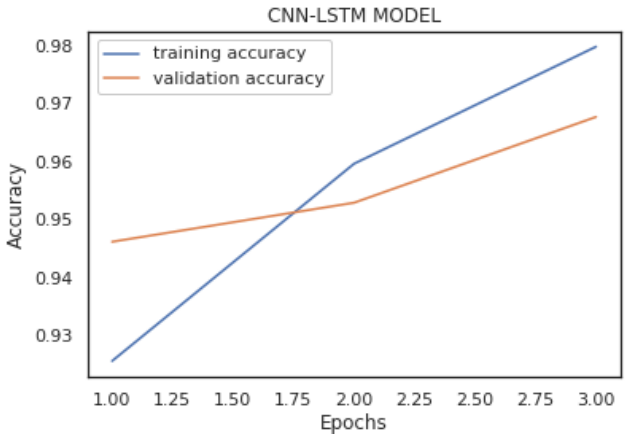

Figure 3: Accuracy for CNN-LSTM.

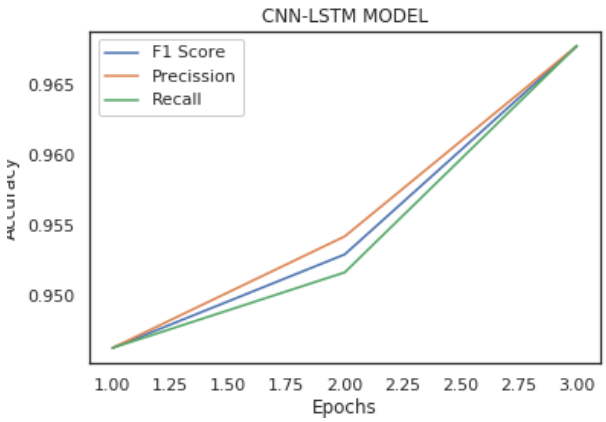

Figure 4: Result for CNN-LSTM Metrics.

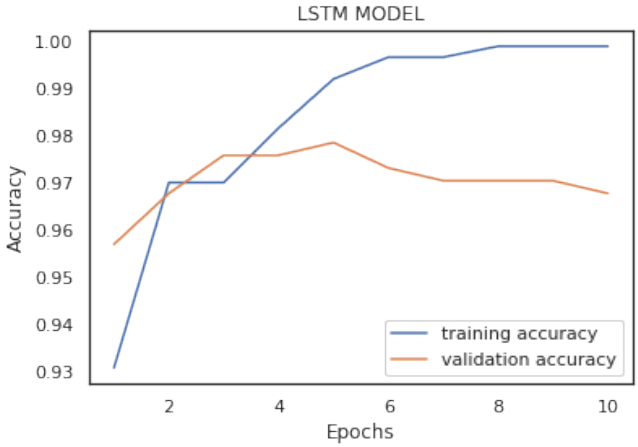

Figure 5: Accuracy for LSTM Metrics.

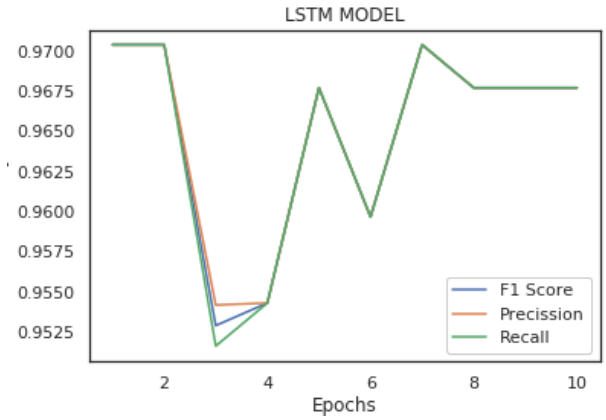

Figure 6: Result for LSTM Metrics.

VI. Conclusion and Future Work

We have provided a large number of annotated personal stories of sexual harassment. To change these patterns and to change social tolerance, it is essential to analyse and identify social patterns of sexual harassment. In this paper

we used three neural models: CNN, LSTM, CNN-LSTM, and BiLSTM. The future extension of this work can focus on hyper-parameter tuning and more comparisons to other models. We learned that data collection was one of the main challenges of this research. Therefore, we would like to develop a customized web scrapper to generate data more efficiently and effectively. Due to lack of resources we were unable to train and test our models extensively. The future work of this study can re-implement the introduced models using better hardware and higher epochs.

REFERENCES

[1] L. Manikonda, G. Beigi, S. Kambhampati, and H. Liu, "# metoo through the lens of social media," in *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 2018, pp. 104–110.

[2] L. Gao, A. Kuppersmith, and R. Huang, "Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, Nov. 2017.

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

[4] C. Van Hee, E. Lefever, and V. Hoste, "We usually don't like going to the dentist: Using common sense to detect irony on twitter," *Computational Linguistics*, vol. 44, no. 4, pp. 793–832, 2018.

[5] G. Szarvas, V. Vincze, R. Farkas, G. Móra, and I. Gurevych, "Cross-genre and cross-domain detection of semantic uncertainty," *Computational Linguistics*, vol. 38, no. 2, pp. 335–367, 2012.

[6] M. Riedl and C. Biemann, "Using semantics for granularities of tokenization," *Computational Linguistics*, vol. 44, no. 3, pp. 483–524, 2018.

[7] A. Khatua, E. Cambria, and A. Khatua, "Sounds of silence breakers: exploring sexual violence on twitter," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 397–400.

[8] S. Merity, N. S. Keskar, and R. Socher, "Regularizing and optimizing lstm language models," *arXiv preprint arXiv:1708.02182*, 2017.

[9] D. I. H. Farías, V. Patti, and P. Rosso, "Irony detection in twitter: The role of affective content," *ACM Transactions on Internet Technology (TOIT)*, vol. 16, no. 3, pp. 1–24, 2016.