CMSI 537 Natural Language Processing, Final Projects
17 December 2020; Los Angeles, CA

Identifying Domestic Violence in Social Media Posts

Lucille Njoo
Loyola Marymount University
Computer Science Department
Los Angeles, United States
lnjoo@lion.lmu.edu

Manny Barreto
Loyola Marymount University
Computer Science Department
Los Angeles, United States
mbarre13@lion.lmu.edu

Cooper LaRhette
Loyola Marymount University
Computer Science Department
Los Angeles, United States
clarhett@lion.lmu.edu

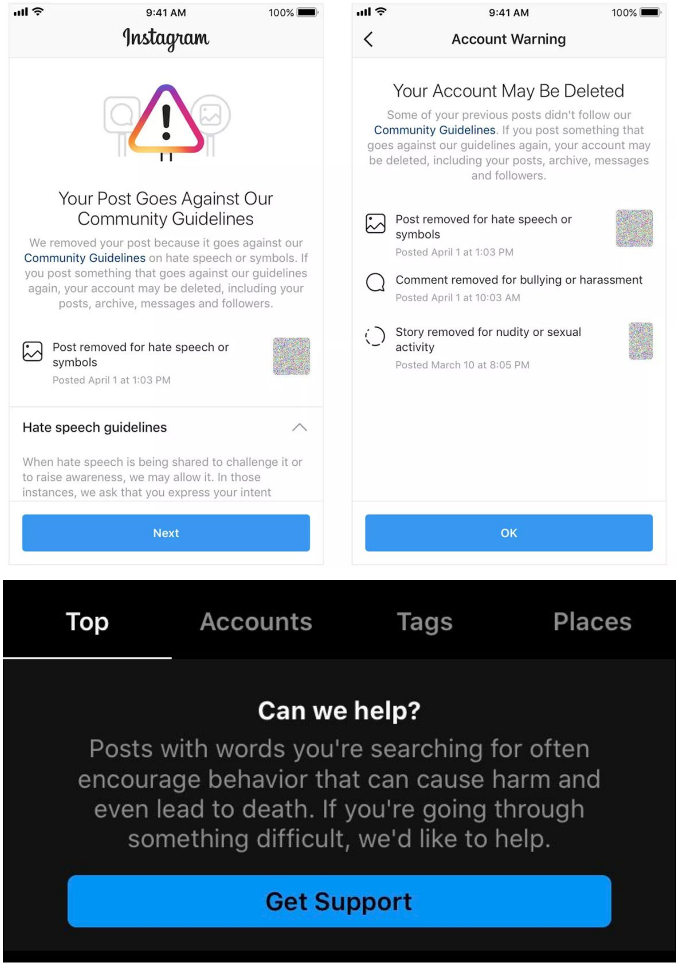
I. Abstract

Domestic violence (DV) is a serious threat to the health and safety of victims, and it can be lifesaving to identify active situations of DV in a timely manner [1] [2] [3] [4]. Victims of DV often turn to social media to vent and seek support, raising the opportunity for applications to identify potential victims of DV and provide them with professional resources. This project develops text classification models using a variety of architectures to identify critical DV situations in social media posts in order to provide assistance to potential victims of DV. Though there is very little existing work in this specific application domain, our project builds off of Subramani et al. [5] by using a different, larger dataset of Reddit posts published by Schrading [6] and supplements this data using synonym-based data augmentation. Overall, our models achieved lower precision, recall, f1-scores, and overall accuracies than Subramani's models with the same architectures, with the exception of our baseline logistic regression classifier, which had comparable performance. This decrease in performance is likely largely due to limitations in computational power and time that prevented us from tuning the models sufficiently; in addition, we also suggest that Subramani's models were overfitted on their very small dataset and that the increased noise in our dataset contributed to our lower accuracy.

II. Introduction

Domestic violence (DV), broadly defined as any kind of physically, sexually, or emotionally abusive behavior in the home, has devastating effects on victims and can seriously threaten their mental and physical health, their safety, and sometimes even their lives [1] [2] [3] [4]. Identifying active situations of DV that pose real risk to victims can be lifesaving when done in a timely manner, but victims of DV are often hesitant to make formal reports or seek help from external services, causing many cases of DV to remain undetected [7]. However, victims often turn to social media to vent or seek empathy, support, or advice [5]. This raises the opportunity for social media applications to detect when a user writes a post that indicates that they may be experiencing DV.

Social media applications already implement a similar response to other kinds of posts that are of noteworthy concern, and many previous studies have trained text classifiers on social media data in order to flag such posts. For example, previous works have built models for detecting hate speech with logistic regression classifiers [8] as well as with more sophisticated models like RNNs (recurrent neural networks) [9] and state-of-the-art embeddings like BERT (Bidirectional Encoder Representations) [10]. Other works have developed models to identify suicidal ideation in social media posts using RNNs, LSTMs (long short-term memory neural networks), and CNNs (convolutional neural networks) [11]. Popular social media applications like Instagram have already integrated some such models as application features; users are informed when their posts are detected as containing hate speech, and they are offered resources when their search queries indicate suicidal or self-harming intents, as shown in Figure 1. Since DV affects the health and safety of users, applications can similarly benefit from detecting DV in users' posts, taking advantage of the fact that victims frequently turn to social media for support [7]. Applications would then be able to offer resources directly to users, which could potentially reach more victims than formal services while preserving their sense of privacy. To this end, our project aims to train models that can detect situations of active DV in social media posts.



Your Post Goes Against Our Community Guidelines

We removed your post because it goes against our Community Guidelines on hate speech or symbols. If you post something that goes against our guidelines again, your account may be deleted, including your posts, archive, messages and followers.

Post removed for hate speech or symbols
Posted April 1 at 1:03 PM

When hate speech is being shared to challenge it or to raise awareness, we may allow it. In those instances, we ask that you express your intent

Your Account May Be Deleted

Some of your previous posts didn't follow our Community Guidelines. If you post something that goes against our guidelines again, your account may be deleted, including your posts, archive, messages and followers.

Post removed for hate speech or symbols
Posted April 1 at 1:03 PM

Comment removed for bullying or harassment
Posted April 1 at 10:03 AM

Story removed for nudity or sexual activity
Posted March 10 at 8:05 PM

Can we help?

Posts with words you're searching for often encourage behavior that can cause harm and even lead to death. If you're going through something difficult, we'd like to help.

Fig. 1. Examples of Instagram's messages when a user writes a post that contains hate speech (top) and when a user enters a search query for a concerning topic, such as suicide or self harm (bottom).

Our project builds off of two previous studies that developed text classifiers on DV-related datasets. Subramani et al.'s work [5], the primary basis of our current project, constructed a labeled dataset of 1,654 Facebook posts extracted from DV support groups, categorizing them under 5 labels describing the type of post. Using this dataset, the researchers implemented 4 machine learning approaches, including logistic regression, as a baseline, then compared their performances to 5 deep learning models: CNNs, RNNs, LSTMs, GRUs, and BLSTMs. In addition, they experimented with a variety of input embeddings on all the models, concluding that GRUs with GloVe embeddings performed the best. The second foundational work for the current project is Schrading's analysis of DV in Twitter and Reddit posts [6]. Schrading extracted tweets with the hashtags #WhyIStayed and #WhyILeft, as well as Reddit posts from nine different subreddits that he labeled as either "abuse" or "non-abuse." For both datasets, Schrading analyzed lexical statistics, such as the most frequently occurring n-grams, common verbs, and subject-verb-object structures. He then trained several machine learning classifiers on both the tweets and the Reddit data and evaluated the different models' efficacies, then utilized these models' insights to better understand the dynamics of abusive relationships.

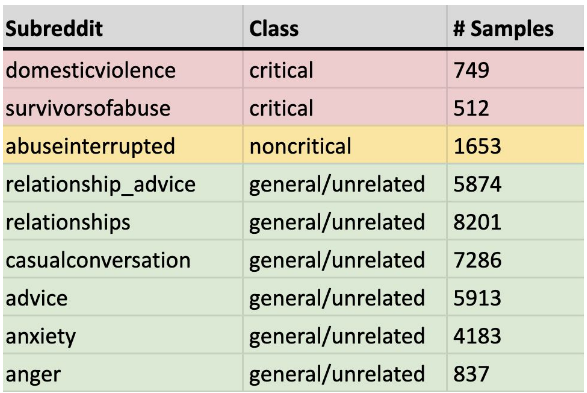
Our approach builds off of these two studies by implementing several of the same models as Subramani et al., but on a larger and noisier new dataset of Reddit posts published by Schrading, with modified labels and supplemented by data augmentation; in doing so, we aim to not just produce models that can accurately detect critical DV situations, but also observe how the models perform on a new dataset and compare their performance to the original Subramani et al. study.

III. Methodology

A. Overview

To replicate the work done by Subramani et al. [5] on Schrading's Reddit dataset [6], we first preprocess the data to label each post with our custom classes, "critical," "noncritical," and "general/unrelated." Like Subramani et al. did, we begin with a simple machine learning model as a baseline. We then implement three of the same neural network architectures that the original work used (BLSTM, RNN, and GRU) and train and evaluate each of these models on the original dataset. Then, we augment the dataset by producing new samples with random words replaced with their synonyms, and we train and evaluate all the models on this augmented dataset as well. We compare our results on both the original and the augmented dataset to Subramani et al.'s original study. In doing so, we can draw conclusions about the effects of using a larger but noisier dataset than Subramani et al.'s original Facebook dataset.

TABLE I. Subreddit Labels and Sample Counts



Subreddit	Class	# Samples
domesticviolence	critical	749
survivorsofabuse	critical	512
abuseinterrupted	noncritical	1653
relationship_advice	general/unrelated	5874
relationships	general/unrelated	8201
casualconversation	general/unrelated	7286
advice	general/unrelated	5913
anxiety	general/unrelated	4183
anger	general/unrelated	837

B. Dataset and Preprocessing

Our project uses Schrading's Reddit dataset of posts from 9 different subreddits (forums dedicated to certain topics). In [6], Schrading categorized these subreddits as either "abuse" or "non-abuse" based on the subreddit's title, and he strategically selected the "non-abuse" subreddits as control groups to help models identify features unique to DV-related posts; many victims of DV may express negative emotion, discuss relationships, or ask for advice, but mentions of emotions, relationships, or advice on their own do not make a given post DV-related [6].

We preprocessed Schrading's dataset with three classes rather than just two: "critical," indicating an active DV situation in which someone may be in immediate danger; "noncritical," indicating posts that are about DV or tangentially related to DV but not about an at-risk situation; and "general/unrelated" for all other posts. Notably, this is also slightly different from the five classes that Subramani et al. used, which were "personal story", "fund raising", "awareness", "empathy", and "general" [5]. By reducing the number of classes from five to three, we hoped to potentially improve performance; furthermore, these three classes are more practical for real-world applications that realistically only need to detect critical DV situations rather than any references to DV at all.

Thus, all of Schrading's "non-abuse" subreddits were labeled as "general/unrelated," but we chose to split the "abuse" subreddits into "critical" and "noncritical." To do so, we manually read through posts on each of the pages to determine whether they would potentially be flagged for an active DV situation; *domesticviolence* and *survivorsofabuse* fell under "critical" because they primarily contained victims' personal stories and their requests for advice and support about current abusers. In contrast, *abuseinterrupted* contained mostly links to articles and general statements about abuse, so it fell under "noncritical."

The total of roughly 30,000 samples is significantly larger than Subramani et al.'s dataset of only about a thousand samples, and also includes fewer classes, which we predicted would contribute to higher accuracies. However, there are some risks in the data itself. First, the classes are heavily unbalanced, with only about a thousand each of "critical" and "noncritical" and the remaining roughly 30,000 as "general/unrelated." Furthermore, the samples are not individually hand-labeled; we did not read all 30,000 posts, and there might be outliers in any of the classes that could throw models off. For example, someone might have posted about a real abusive relationship in the relationship_advice subreddit, but their post would be lumped under the "general/unrelated" label since we aggregate entire subreddits under labels rather than individual posts.

We re-labeled the Scrading dataset as described above and split the samples into an 80% training dataset and 20% held-out testing dataset. Each of the posts were then tokenized into unigrams for bag-of-words features by splitting the strings on spaces.

C. Models

Like Subramani et al. did, we begin with a simple machine learning model as a baseline. Subramani et al. used four machine learning models: Support Vector Machine, Decision Trees, Random Forests, and Logistic Regression [5]. We chose to implement a logistic regression classifier as our baseline.

We then implemented three neural network architectures: Bidirectional Long Short-Term Memory (BLSTM), Gated Recurrent Unit (GRU), and Recurrent Neural Network (RNN). These were built using the PyTorch library's neural network module. We based these models' hyperparameters largely on those used by Subramani et al. for the same architectures (see Table 1). For all three models, we used Stochastic Gradient Descent as the optimizer and Cross Entropy Loss as the loss function.

TABLE II. MODELS AND HYPERPARAMETERS

Model	Hidden Size	Learning Rate	Layers	Epochs	Batch Size
LR	N/A	N/A	N/A	N/A	N/A
RNN	50	0.01	1	20	50
BLSTM	25	0.01	2	15	40
GRU	50	0.01	1	20	50

D. Model Evaluation

We evaluate all our models using precision, recall, f1-score, and accuracy and compare these metrics to the results from Subramani et al.'s original study. All models are evaluated on the same 20% held-out test set. Notably, our project slightly diverges from the original study at this step as well; Subramani et al. used k-fold cross-validation and averaged their results on the *k* different testing sets for their final performance metrics [5]. This was likely necessary in their study because their dataset was so small that it would be easy for complex models to overfit; we opted not to use the k-fold cross validation in our project because our dataset is much larger and noisier, which inherently prevents overfitting.

E. Data Augmentation

Since there was a heavy class bias towards the “general/unrelated” label in the Schrading dataset, we generated additional data for the two underrepresented classes to reduce our models' bias towards the dominant class. Since the original dataset had about 30,000 samples of the "general/unrelated" class but only about 2,000 of the first two classes combined, we generated approximately 20,000 new samples, split evenly between the two DV-related classes, “critical” and “noncritical”.

To do this, we used nlpaug, an open source library developed for NLP data augmentation. For each of the underrepresented classes, we filtered out that class from the original Reddit data and used this as a seed dataset from which we could randomly pick a sample to generate from. We then used nlpaug's synonym substitution method based on the WordNet lexical database to generate a new sample that is almost the same as the original, but with some words replaced with synonyms. We repeated this iteratively until we had generated 10,000 samples, and then did the same for the other underrepresented class.

The augmented dataset has approximately 52,000 total data samples to train and test on. As before, we randomly split this dataset into 80% training and 20% testing sets so that all models can be trained, evaluated, and compared on the same data.

IV. RESULTS

A. Using Original Reddit Data

On the unaugmented Reddit data, our models appear to do comparably well with Subramani et al.'s models in terms of accuracy. However, the high accuracy paired with the low precision, recall, and f1-scores, especially for our RNN and GRU, was evidence that the models were only achieving high accuracies because they were almost always predicting the overrepresented class, "general/unrelated". The exception is the logistic regression model, which had high accuracies as well as precision, recall, and f1-score. This class imbalance is why the data augmentation was necessary; without it, the neural network models would always predict that a post is not DV-related, which defeats the purpose of having the model in the first place.

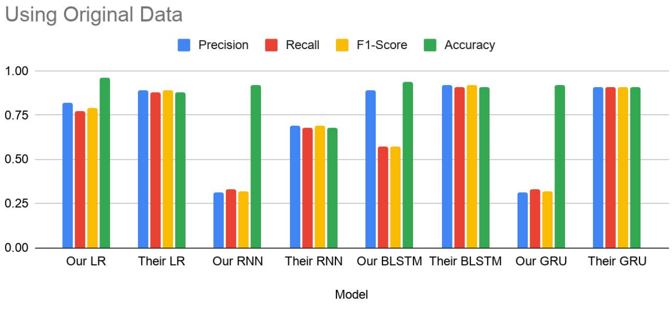


Fig. 2. Comparing performance of our models trained on original data to Subramani et al.'s models

B. Using Augmented Reddit Data

On the augmented dataset, our models overall did worse than the same models did on Subramani et al.'s dataset, but they were less biased towards the "general/unrelated" class. The precision, recall, and f1-scores were still low, but the accuracies were no longer drastically higher, which aligns with the fact that the models were no longer solely predicting the overrepresented class. With more balanced data, these decreased accuracies are better measures of the models' true accuracies, showing that the models can predict all 3 classes, but not very reliably yet. The exception is again our baseline logistic regression model, which achieved very high accuracies as well as precision, recall, and f1-scores, outperforming not just our neural networks but all of the original paper's models.

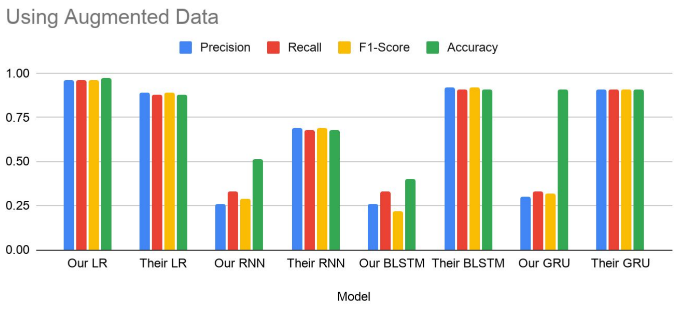


Fig. 3. Comparing performance of our models trained on the augmented dataset to Subramani et al.'s models

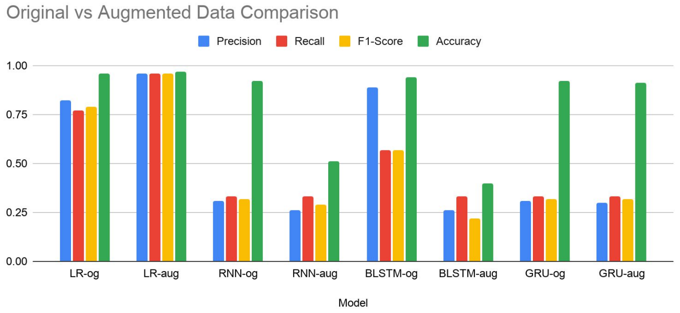


Fig. 4. Comparing our models' performance on the original vs. the augmented datasets

V. Analysis and Discussion

Looking at our baseline logistic regression model alone, we can confirm that our larger dataset did help to increase accuracy, and augmenting the data further increased all metrics. Our logistic regression model outperformed all of Subramani et al.'s models with an overall accuracy of 97%, demonstrating the utility of having a larger dataset and fewer classes, as we expected.

Our results for all models on the original dataset also demonstrates the dangers of unbalanced data. The models appeared to be performing well in terms of accuracy, but they actually were incapable of recognizing the two underrepresented classes entirely. With unbalanced data, models become very good at recognizing classes for which there are plenty of samples, but very poor at recognizing classes that lack quality data.

Though we expected the three neural network architectures to perform better than our logistic regression baseline, the opposite was true -- on both the unaugmented Reddit data and the augmented Reddit data, our logistic regression model did better than all our neural network models. The logistic regression model's success shows that the classes of our new dataset are distinct enough to be distinguished with a simple model, and architecturally, there should be no reason why a larger model with a more complex, nonlinear function would be unable to fit to the same data. Thus, the neural networks' poor performances can likely be largely attributed to insufficient tuning of hyperparameters. We spent several days tuning our models, but were only able to make marginal improvements due to our computational limitations. The neural network models take a very long time to train, so we were limited in how many times we could try running them.

This tradeoff between speed and model complexity is not always worth it. In this project, our logistic regression model performed very well, and likely would be accurate enough to use in real social media applications. Though the neural network architectures could potentially be even more accurate than the logistic regression models, they took exponentially longer to train, both in terms of man-hours spent manually tuning hyperparameters as well as actual time spent training, even on Google's Cloud Platform. They also required much more memory both to train and to store, and Google Colab was unable to handle some of the larger variants of these models because of its memory limits.

The poor performance of our neural network models on a different dataset could also be interpreted to indicate that Subramani et al.'s models were overfitted to their very small dataset. With their dataset of only a thousand samples across five classes, it is likely for high-capacity models like neural networks to overfit, and the fact that we used the same architectures and hyperparameters but received drastically different results indicates that perhaps the Subramani et al. model would not have been able to generalize to different datasets; the language of Facebook posts might be significantly different from the language of Reddit posts, causing our models to need very different hyperparameters than Subramani et al.'s. This lack of generalizability is a common problem in many neural network architectures, and Subramani et al.'s original publication discussed this as well; neural networks can be very brittle, and their hyperparameters as well as the optimal architecture are highly dependent on the dataset and application task [5]. Given neural networks' propensity for overfitting, and given the fact that the topic of DV in social media is not one that has been widely explored, and not one for which there are widely available datasets from different sources, this might indicate that a simpler model like our logistic regression classifier might do better than more complex models because they are less likely to overfit to a single dataset.

VI. Future Work

For future research, we hope to be able to spend more time tuning the three neural network models because as higher-capacity models, they should be able to match if not improve upon the baseline logistic regression model. In addition, it would be interesting to condense the classes from three to just two, "DV-related" and "unrelated", similar to Schrading's original labeling of his Reddit data. This might increase accuracy and improve performance on our metrics, but it might be less useful in real-world applications, because not all mentions of DV in social media posts are cause for concern; some are awareness-related or retrospective, neither of which should be flagged for being a critical situation. In addition, it would be interesting to supplement this data with posts from another social media source. For example, Twitter data from the #whyistayed and #whyileft hashtags, a 2014 trend in support of DV survivors, might be helpful in creating models that are more generalizable. Finally, there may be merit for future research in experimenting with the models' features. For example, Schrading's Reddit dataset contained a VADER sentiment score for each post that we did not incorporate into our models' features; adding this could potentially improve performance. It may also be worthwhile to investigate how other NLP tasks could be used in concert with the token features to improve performance in this text classification problem, such as semantic role labeling, especially because acts of DV are often characterized by particular people, relationships, and actions.

VII. References

[1] A. Levendosky, S. Graham-Bermann, Parenting in Battered Women: The Effects of Domestic Violence on Women and Their Children, 2001, Journal of Family Violence, Vol. 16, No. 2.

[2] S. Walby, 2004, *The Cost of Domestic Violence.* Available: https://www.researchgate.net/publication/266474561_The_Cost_of_Domestic_Violence

[3] C. Garcia-Moreno, A. Guedes, and W. Knerr, "Understanding and addressing violence against women." World Health Organization, Geneva, Switzerland, 2012.

[4] C. Garcia-Moreno, C. Pallitto, K. Devries, H. Stöckl, and C. Watts, "Global and regional estimates of violence against women: prevalence and health effects of intimate partner violence and non-partner sexual violence."

[5] S. Subramani, S. Michalska, H. Wang, J. Du, Y. Zhang and H. Shakeel, "Deep Learning for Multi-Class Identification From Domestic Violence Online Posts," in IEEE Access, vol. 7, pp. 46210-46224, 2019. Available: doi:10.1109/ACCESS.2019.2908827.

[6] J. N. Schrading, "Analyzing Domestic Abuse using Natural Language Processing on Social Media Data," M.S. Thesis, Dept. of Computer Engineering, Rochester Institute of Technology, Rochester, (NY), 2015. Available: https://scholarworks.rit.edu/cgi/viewcontent.cgi?article=9949&context=theses

[7] M. A. Evans and G. S. Feder, ‘‘Help-seeking amongst women survivors of domestic violence: A qualitative study of pathways towards formal and informal support,’’ Health Expectations, vol. 19, no. 1, pp. 62–73, 2016.

[8] T. Davidson, D. Bhattacharya, and I. Weber, "Racial Bias in Hate Speech and Abusive Language Detection Datasets", Third Abusive Language Workshop at the Annual Meeting for the Association for Computational Linguistics 2019,May 2019. Available: https://arxiv.org/pdf/1905.12516.pdf.

[9] G.K. Pitsilis, H. Ramampiaro, and H. Langseth, "Effective hate-speech detection in Twitter data using recurrent neural networks," Appl Intell vol. 48, pp. 4730–4742, July 2018. Available: https://doi.org/10.1007/s10489-018-1242-y

[10] M. Mozafar, R. Farahbakhsh, and N. Crespi, "A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media," Studies in Computational Intelligence, vol. 88, 2019. Available: https://doi.org/10.1007/978-3-030-36687-2_77.

[11] S. Ramit, et al. "Exploring and learning suicidal ideation connotations on social media with deep learning." Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. 2018.