NLP Final Project: Identifying Domestic Violence in Social Media Posts
Milestone Report

Lucille Njoo, Manny Barreto, Cooper LaRhette

I. Revised Objective / Hypothesis

The objective of our research project is to implement several variations of a domestic violence text classifier based on the architectures used by Subramani et al. in [1], but using a different, larger dataset of Reddit posts taken from Schrading [2] rather than their original dataset of Facebook posts. By applying Subramani's work to a new dataset, we hope to be able to compare the performance of different neural network architectures and juxtapose our models' results with the original paper's to draw conclusions about the differences in the two datasets and their efficacies.

II. Literature Review

Text classifiers are often trained on social media data in order to flag posts that are of noteworthy concern; several previous works have built models for detecting hate speech with logistic regression classifiers [3] as well as with more sophisticated models like RNNs [4] and state-of-the-art embeddings like BERT [5]. Besides hate speech, domestic violence (DV) is another topic of concern for detection in social media posts, especially since victims frequently turn to social media for support [1]. Our project is inspired by two studies that built text classifiers on DV-related datasets. Subramani et al.'s work [1], the primary basis of our current proposal, 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 with all the models, concluding that GRUs with GloVe embeddings performed the best. The second major inspiration for our project is Schrading's analysis of DV in Twitter and Reddit posts. Schrading extracted tweets with the hashtags #WhyIStayed and #WhyILeft, as well as Reddit posts from seven 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.

III. Revised Proposed Project Approach

First, we aim to replicate the work done by Subramani et al. [1] on a new dataset of Reddit posts published by Schrading [2]. Though Subramani used 5 classes, we aim to categorize posts as "critical," "noncritical," and "general/unrelated." By reducing the number of classes from 5 to 3, we can potentially improve performance; furthermore, these 3 classes are more practical for real-world applications that might need to detect only critical DV rather than any references to DV at all. Like Subramani et al. did, we plan to begin with a simple machine learning model as a baseline. Though the original study implemented 4 different models, for the sake of time, we will only be implementing a logistic regression classifier. We then will move on to implementing 3 of the same neural network architectures that the original work used: LSTM, RNN, and GRU classifiers. We can then evaluate all our models using precision, recall, f1-score, and accuracy to see how our results on the new dataset compare with Subramani et al.'s original study. Finally, building off of Subramani et al. and Schrading's works by drawing from other successful social media classifiers implemented in [4] and [5], we aim to experiment with different pre-trained word embeddings like word2vec and BERT. In doing so, we hope to be able to draw conclusions about not just the efficacies of using these embeddings compared to the models without them, but also about the benefits or drawbacks of having a larger but noisier dataset than Subramani et al.'s original Facebook dataset.

IV. Revised Tentative Schedule

- *Week 1-2 (11/2 – 11/13):* Load and preprocess the Schrading Reddit dataset (all three team members). We ran into a lot of data loading issues, so this took much longer than expected.
- *Week 3 (11/16 – 11/20):* Build a baseline logistic regression classifier (Lucille) and two neural network models: an LSTM (Lucille and Manny) and an RNN (Cooper); compare their performances.
- *Week 4 (11/23 – 11/27):* Milestone Week; complete our 4th model, a GRU (Manny, Lucille); start trying the models with a couple of different embeddings (word2vec, Manny; BERT, Lucille and Cooper).

- *Week 5 (11/30 – 12/4):* Finish implementing the models with pre-trained embeddings (all three team members), then compare the performances of the models with the new embeddings to both our baseline models and the original paper's.
- *Week 6 (12/7 – 12/11):* In-class presentations; clean up code and write our final report (all three team members).
- *Week 7 (12/14 – 12/18):* Final report and code due; make final tweaks and submit report (all three team members).

V. Dataset and Data Preprocessing

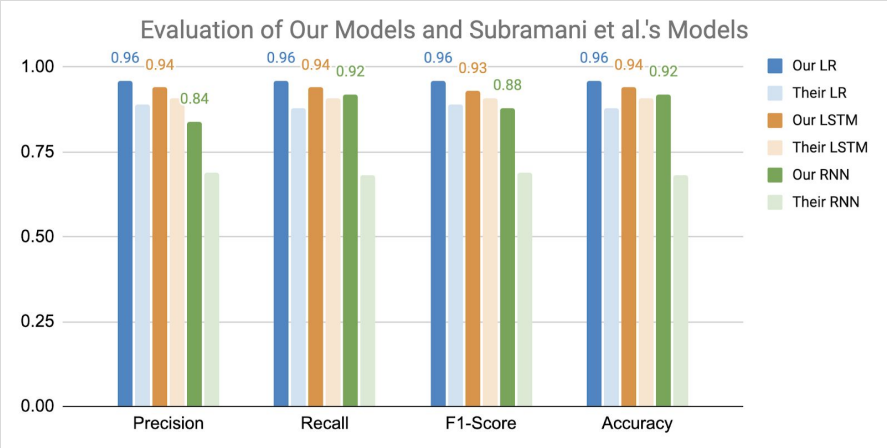
Our project uses Schrading's Reddit dataset of posts from 9 different subreddits (forums dedicated to certain topics); in [2], Schrading categorized these subreddits as either "abuse" or "non-abuse" based on the subreddit's title, and he strategically selected the "non-abuse" subreddits to help models identify features unique to DV-related posts; many victims of DV may express 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 [2].

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

We preprocessed Schrading's dataset with 3 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. 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 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' stories/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 predict will contribute to higher accuracies. However, there are some risks in the data itself; the classes are unbalanced, with only about a thousand each of "critical" and "noncritical" and the rest as "general/unrelated." Furthermore, it is not hand-labeled; our team has 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 post about an abusive relationship in the *relationship_advice* subreddit, but their post would be labeled as "general/unrelated."

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.

VI. Preliminary Results



Though we have only implemented three models so far, our preliminary results show very high accuracies, even on the most basic logistic regression model. The logistic regression model achieves a 96% accuracy on the held-out test data, with average precision, recall, and f1-scores of 96% as well. Surprisingly, the LSTM model is less accurate, and it achieves only about a 94% accuracy. Our RNN model performs even worse, with a 92% accuracy. The relatively poor performance of our LSTM and RNN is likely because we have not had the opportunity to train them with as many dimensions and epochs as we had planned; however, even with the small number of hidden dimensions and epochs we are currently using, our models so far perform better than Subramani et al.'s models; their LSTM achieved about 91% accuracy, and their RNN only reached a 68% accuracy, so it makes sense that our RNN's performance is currently the worst. We think our models have so far demonstrated better accuracies because of the significant increase in the amount of data; our Reddit dataset is an entire order of magnitude larger than Subramani et al.'s dataset. We remain cautious of the fact that these posts were not hand-labeled and that they may contain noise and outliers rather than falling neatly into one of the three boxes. However, this might actually contribute to better testing

accuracies since imperfect data can sometimes prevent overfitting. Our next steps are to train our LSTM and RNN models for more epochs, then compare it with a GRU model before adding pre-trained word embeddings like BERT.

VII. References

[1] 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.

[2] 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

[3] 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.

[4] 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

[5] 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.