NLP Final Project: Identifying Domestic Violence in Social Media Posts

Lucille Njoo, Manny Barreto, Cooper LaRhette

I. Objective / Hypothesis

The objective of our research project is to reimplement a slightly simplified version of the domestic violence text classifier created by Subramani et al. in [1], and to build off of their work by applying augmenting the original dataset with new datasets [2] and comparing the performance of different neural network architectures.

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 detecting 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 samples extracted from Facebook DV support groups, using 5 categories that describe the type DV post. Using this dataset, the researchers then 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 fell into the categories of "abuse" and "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, and used these models to better understand the dynamics of abusive relationships.

III. Proposed Project Approach

First, we aim to replicate the work done by Subramani et al. [1], but with their dataset's labels condensed from the original five to just three: critical, non-critical, and general/unrelated; this not only may potentially improve performance given the small dataset size, but is a more practical categorization for real-world applications. Like Subramani et al.'s study, 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 can then move on to implementing 3 of the same neural network architectures that the original work used: LSTM, RNN, and GRU classifiers. Drawing from other successful social media classifiers in [4] and [5], we aim to experiment with different pre-trained word embeddings like word2vec and BERT. Our next step would be expanding our dataset, first by relabeling the Reddit data gathered by Schrading [2] and augmenting it to our original dataset, then by utilizing the nlpaug data augmentation library to further expand the dataset. We can then evaluate all our models on the expanded dataset using precision, recall, f1-score, and accuracy to see if increasing the dataset size improves performance. By comparing these performances to the original paper's results, we hope to be able to draw conclusions about how reducing the number of labels and increasing the dataset size affects the efficacy of these models.

IV. Tentative Schedule

- *Week 1 (11/2 – 11/6):* Preprocess the Subramani et al. dataset (Manny, Cooper); build a logistic regression classifier (Lucille).
- *Week 2 (11/9 – 11/13):* Implement LSTM, RNN, and GRU classifiers with the same dataset (one model per team member).
- *Week 3 (11/16 – 11/20):* Experiment with different pre-trained word embeddings across the different models, including word2vec (Cooper) and BERT (Manny, Lucille), and compare their performances.
- *Week 4 (11/23 – 11/27):* Milestone Week; by the end of this week, we plan to have completed our 4 models with a couple of different embeddings, and to compare our results with the original paper's. We plan to begin augmenting the Subramani et al. dataset with Schrading's Reddit dataset (Lucille) and evaluating the models on the larger combined dataset (Manny, Cooper).
- *Week 5 (11/30 – 12/4):* Implement data augmentation using the nlpaug library to further enlarge the dataset, and compare performances; prepare our presentation (all three team members).
- *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. 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.