Detecting Domestic Violence in Social Media Posts

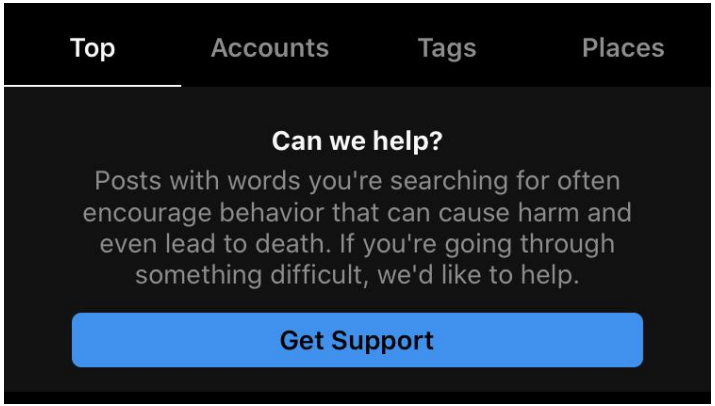
Lucille Njoo, Manny Barreto, Cooper LaRhette

Natural Language Processing Fall 2020

(mild TW)

Motivation & Background

- Identifying **concerning social media posts** has practical use
  - **Domestic violence (DV)** is an issue of safety and health
  - Social media applications could benefit from DV classifiers: flag posts of potential victims and provide resources
- This topic **has not been widely explored!**
- Our project is **based on 2 studies:**
  - **Subramani et al.** - lots and lots of models on a small hand-labeled dataset
  - **Schrading** - more detailed/micro-analysis to understand social dynamics of DV



The Goal... Thrice-Revised

- At first, wanted to reimplement Subramani et al.'s study
  - Despite trying for weeks, we **were not able to obtain the data**
- **Pivoted to using Schrading's data** in an effort to improve performance with the increased number of samples
  - Performance was limited by heavy imbalances in the data
- Turned our focus to **data augmentation**
  - Experimenting with how different augmentation techniques could improve performance/reduce bias towards the over-represented class
- **Hypothesis:** A larger dataset, balanced with data augmentation, will improve performance over Subramani et al.'s models.

The Dataset

- **9 subreddits** that we sorted into **3 classes**: critical, noncritical, and unrelated
- **Pros:** much larger than Subramani et al.'s dataset
- **Cons:** classes are unbalanced, posts are not individually labeled by hand
- Tokenized into **unigram/BoW features**, split into **80% training and 20% testing** sets

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

Models

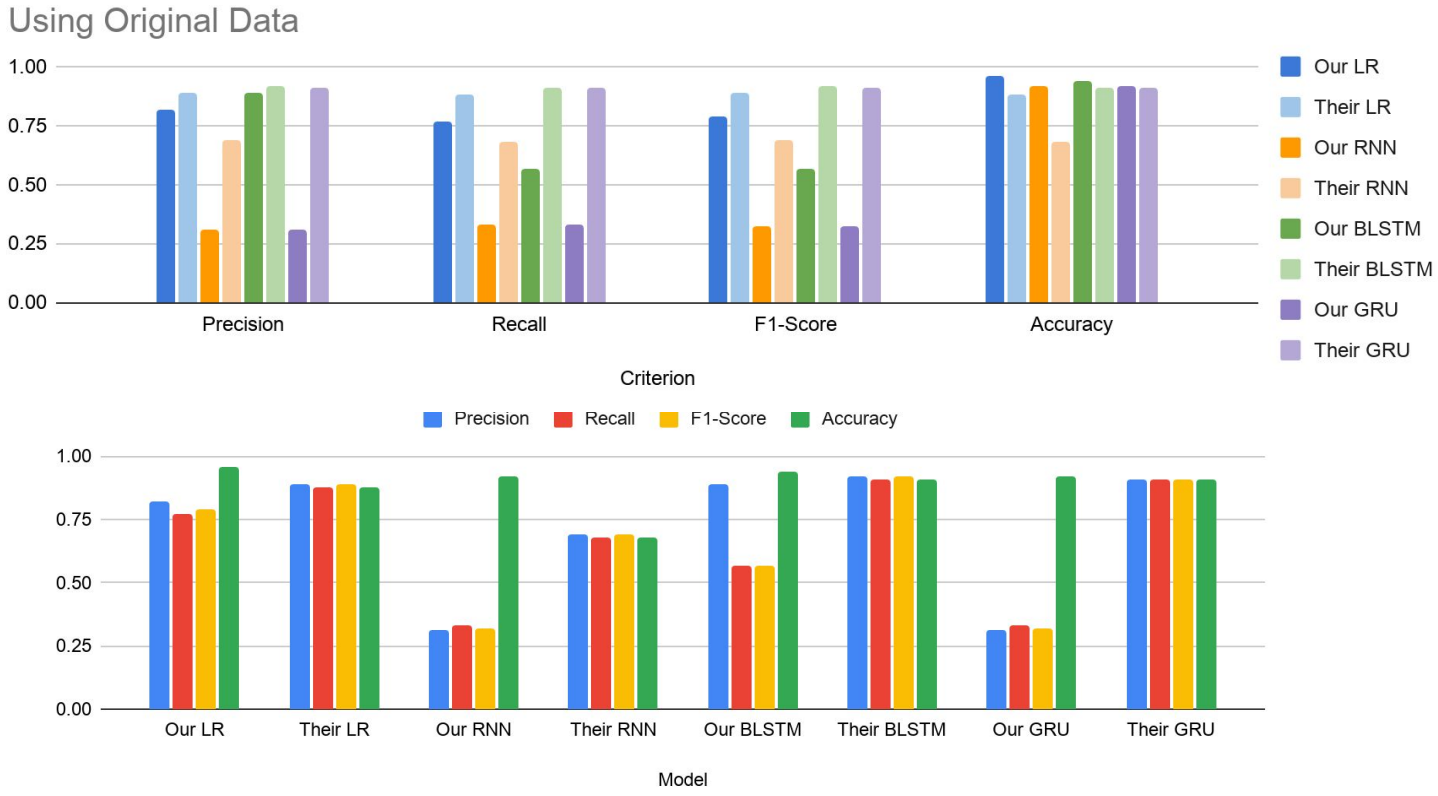
- Used **Logistic Regression** as our baseline
- Chose **3 neural network** architectures
- Evaluated on **precision, recall, f1-score**, and overall **accuracy**

Model	Hidden Size	Learning Rate	Layers	Epochs	Batch Size
Logistic Regression	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

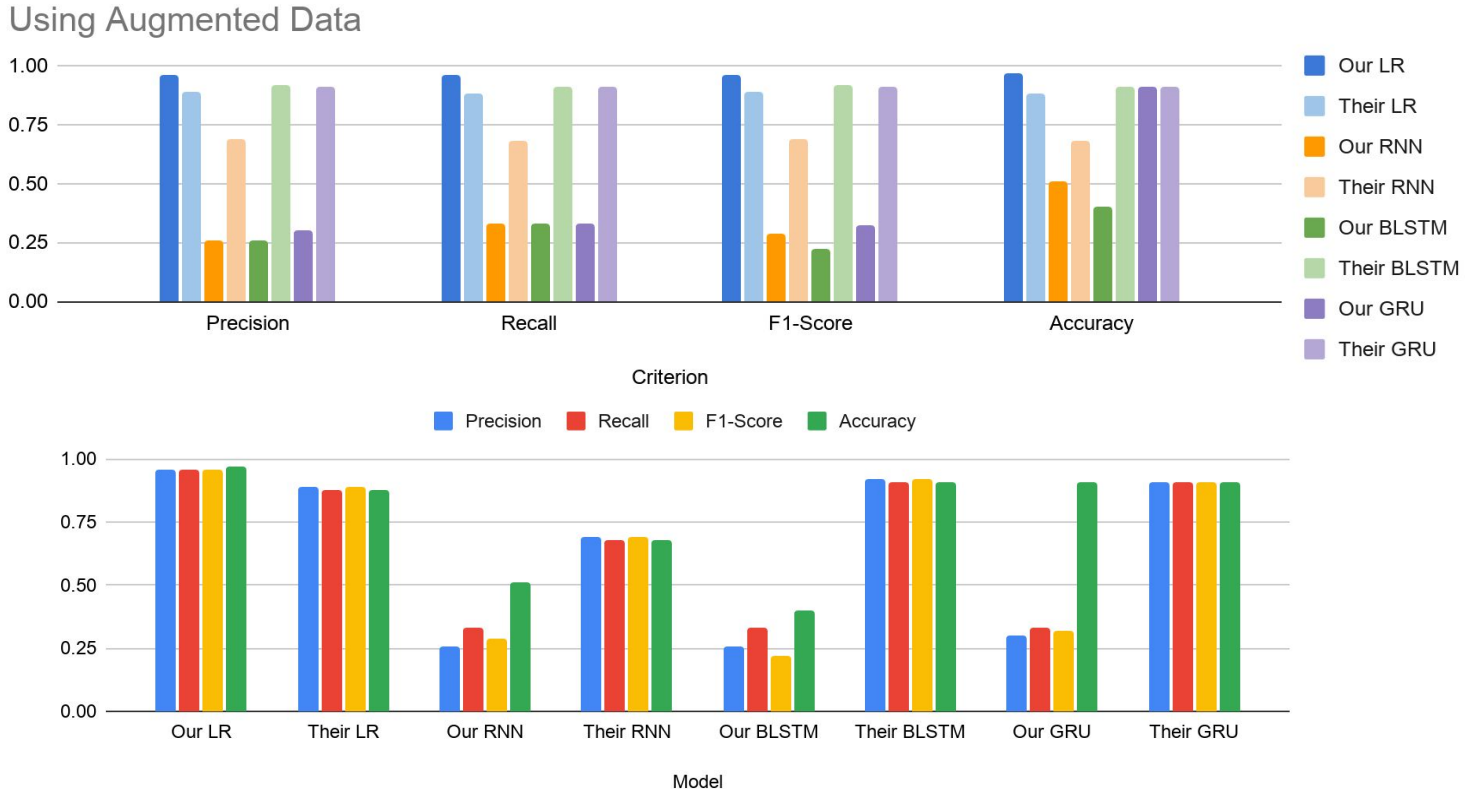
Experimenting with Data Augmentation

- Used the **nlpaug** library for data augmentation
- Tested a variety of models for generating augmented data
  - Static word embeddings using **word2vec**
    - `Original: “The quick brown fox jumps over the lazy dog”`
    - `Augmented Text: “The easy brown fox jumps around the lazy dog”`
  - Contextual word embeddings using **BERT**
    - `Original: “The quick brown fox jumps over the lazy dog”`
    - `Augmented Text: “little quick brown fox jumps over the lazy dog”`
  - Synonym substitution using **WordNet**
    - `Original: “The quick brown fox jumps over the lazy dog”`
    - `Augmented Text: “The speedy brown fox jumps complete the lazy dog”`
- **Synonyms with WordNet** worked best
- **Generated ~20k new samples** to balance the data bias towards class 2

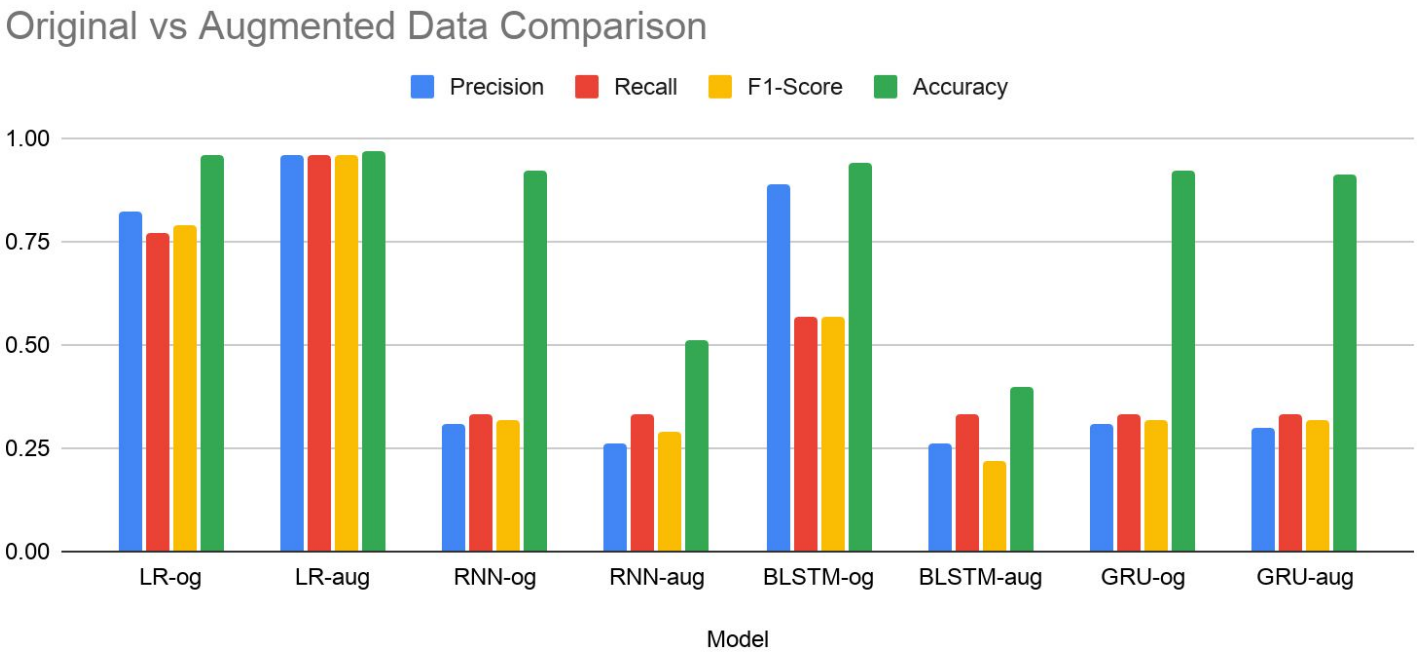
Results: Our Models vs. Subramani et al.



Results: Our Models vs. Subramani et al.



Original vs Augmented Data Comparison

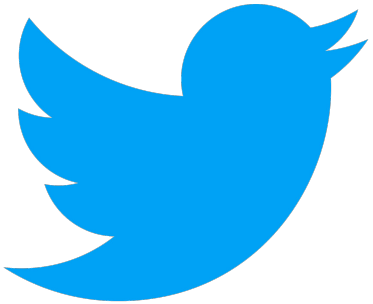


Conclusions

"Bad findings are still findings." --Manny Barreto, 2020

- **Bigger is not always better**; sometimes a simple model does the trick.
  - Our **Logistic Regression model** outperformed all of our neural networks *and* all of the original papers' models!
- **Clean, accurately-labeled data** is crucial.
  - No matter how complex the model, noisy/imbalanced data will have an impact.
- **Data imbalances** can make a model look **misleadingly** good.
  - This often **contributes to bias** in lots of models, which can have bad consequences when the bias is towards a certain skin color, for example
- Original paper's models likely **overfitted on their tiny, curated dataset.**
  - Our models' designs and hyperparameters were heavily based on the paper's, but performed pretty badly on the different dataset.

Potential Next Steps



- **More tuning!**
- Try condensing from 3 to just **2 classes**
  - Models are training as we speak!
- If time allows: Supplement existing DV data with **Twitter data** from the #whyistayed and #whyileft hashtags
  - Perhaps more clearly DV-oriented and distinct from the non-DV-related data, but they are **noisy as well due to misinterpretations and trolls**
- Experiment with **the features**:
  - Use additional features included in Schrading's dataset, such as the **sentiment score**
  - Use **pre-trained embeddings** like word2vec or GloVe

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.

Thank you!

(shoutout to Masao for contributing to our valiant efforts in the battle against the Facebook API)
(curse you, Facebook API!)

Questions?