#MeToo: Neural Detection and Explanation of Language in Personal Abuse Stories

Sweta Karlekar **Mohit Bansal**
UNC Chapel Hill
{swetakar, mbansal}@cs.unc.edu

Abstract

The detection and classification of domestic abuse stories shared online has ever-increasing importance in today's social activism sphere. With massive numbers of stories shared, automatic detection can aggregate stories from around the internet and help push forward the fight against domestic abuse from a social campaign to social change. We develop CNN, LSTM-RNN, and CNN-LSTM neural models to detect domestic abuse stories in the Reddit Domestic Abuse dataset. We achieved 95.8% accuracy in classifying posts as containing abuse stories versus not containing abuse stories, outperforming the current state-of-the-art. More importantly, we next present sentiment-only classification feasibility as well as interpretable and explainable analysis of the neural model's predictions using activation clustering techniques to automatically discover linguistic features.

1 Introduction

Within the past year, the #MeToo[1] hashtag has gained popularity on Facebook, Reddit, Twitter and other social media platforms. This social campaign centers around sharing personal stories about sexual harassment, including domestic and workplace abuse. Many other hashtags have come and gone, including #YesAllWomen, #WhyIStayed, and #ItsNotOkay. However, with each past instance of social outrage, relatively little real-world action was taken towards gender equity and ending gender-based abuse.[2]

The power of natural language processing could serve as one of the missing links between online activism and real change. Deep learning techniques allow us to aggregate, analyze, and summarize vast amounts of data found on social media, becoming a useful tool for spreading awareness. The automatic detection, classification, and interpretation of personal abuse stories can help activist groups educate the public and advocate for social change in a timely fashion.

subreddit	label	entries
abuse-interrupted	abuse	1653
domestic-violence	abuse	749
survivors-of-abuse	abuse	512
casual-conversation	non-abuse	7286
advice	non-abuse	5913
anxiety	non-abuse	4183
anger	non-abuse	837

Table 1: Reddit dataset statistics describing the number of submissions and label collected per subreddit.

In relation to this task, we improve the classification performance of abuse stories via effective CNN, LSTM-RNN, and CNN-LSTM models. Next, we employ activation clustering techniques to explain the features discovered by our neural models. This interpretability technique for automatic feature discovery helps explain the specific language properties that classify certain stories as abuse stories. Further, we demonstrate the limited feasibility of abuse story classification when relying only on sentiment scores.

2 Related Work

Schrading et al. (2015) assembled the Reddit Domestic Abuse dataset, discussed further in Section 3. As one of the first works to address the classification of domestic abuse stories, they used multiple traditional classifiers e.g., Linear SVM, logistic regression, Naïve Bayes, Random Forest, etc. Their highest accuracy of 92.0% was achieved using a Linear SVM (C=1) with N-gram features.

3 Dataset

Reddit is a social media platform that contains a substantially large range of forums called subred-

[1] https://metoomvmt.org
[2] https://www.cnn.com/2017/10/30/health/metoo-legacy/index.html

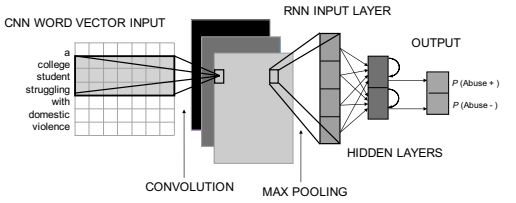

Figure 1: Our CNN-LSTM hybrid neural network.

dits in which users post comments pertaining to a specific topic. These posts are moderated by community volunteers who enforce standard English, respectful behavior, and on-topic discussion. Reddit also allows for lengthy posts. Therefore, Reddit data is ideal for initial ventures into this task.

The Reddit Domestic Abuse dataset (Schrading et al., 2015) is publicly available,[3] and contains submissions labeled "abuse" from *"abuse-interrupted"*, *"domestic-violence"*, and *"survivors-of-abuse"* subreddits. To balance the negative sentiment of "abuse" stores, submissions from *"anger"* and *"anxiety"* subreddits are included as "non-abuse". Submissions from the *"advice"* subreddit are included as "non-abuse" to ensure that classifiers are not just finding help-seeking behavior. The subreddit *"casual-conversation"* is included as "non-abuse" as well. The dataset contains 1336 total cases of "abuse" and 17020 total cases of "non-abuse" (Table 1).

4 Models

CNN: For each input, an embedding and a convolutional layer is applied, followed by a max-pooling layer (Collobert et al., 2011). No pre-trained word embeddings were used. Filter sizes of [3, 4, 5] with 128 filters per filter size were used. The convolution features are then passed to a softmax layer, which outputs probabilities over two classes.

LSTM-RNN: We also adopted an LSTM-RNN (Hochreiter and Schmidhuber, 1997) model with 128 hidden units. The embedding layer was followed by two LSTM hidden layers. The final state is fed to a fully-connected layer and then a softmax layer, which gives the final output probabilities.

CNN-LSTM: We also present a combined CNN-LSTM architecture with complementary strengths of CNNs and LSTM-RNNs (similar to the C-LSTM by Zhou et al. (2015)). Our RNN was laid on top of our CNN model.[4] Please see Figure 1 for more details.

[3] http://nicschrading.com/data/

Model	Accuracy
Schrading et al. (2015)	92.0%
2D-CNN	92.6%
LSTM-RNN	94.5%
CNN-LSTM	**95.8%**

Table 2: Accuracy results on abuse story detection.

5 Results

Table 2 shows classification accuracy of related works as well as our CNN, LSTM-RNN, and CNN-LSTM models. Our best-performing CNN-LSTM model sets the new state-of-the-art standard with an accuracy of 95.8%.[5]

6 Analysis

6.1 Sentiment-Based Classification

Overall sentiment of each submission was calculated by VADER (Gilbert, 2014). More negative posts receive more negative scores and more positive posts receive more positive scores. To ensure that classifier predictions are capturing linguistic characteristics of the stories and are not solely relying on sentiment, we calculated the average sentiment score and standard deviation of each subreddit in the dataset. As shown in Fig. 2, all of the abuse-positive subreddits had negative mean sentiment scores. However, both the "anger" and "anxiety" subreddits (that were part of the non-abuse dataset) had average negative sentiment scores as well. Furthermore, the overall sentiment of each subreddit had large standard deviations. Because of this, no subreddit had a statistically significant difference in sentiment from any other subreddit. Overall, this demonstrates that sentiment alone cannot be used to effectively classify stories as abuse-positive or abuse-negative.

6.2 Activation Clustering

We present activation clustering analysis (following Girshick et al. (2014) and Aubakirova and

[4] A vocabulary size of 10,000 and an Adam-Optimizer (Kingma and Ba, 2015) with learning rate of 1e-4 was used for all models.

[5] CNN, LSTM-RNN, and CNN-LSTM models showed statistically significant difference from each other, calculated using standard deviations, two sample t-tests, and the bootstrap test (Noreen, 1989; Efron and Tibshirani, 1994) where $p < 0.04$. CNN-LSTM and Schrading et al. (2015) showed statistically significant difference calculated based on their reported standard deviations of accuracy.

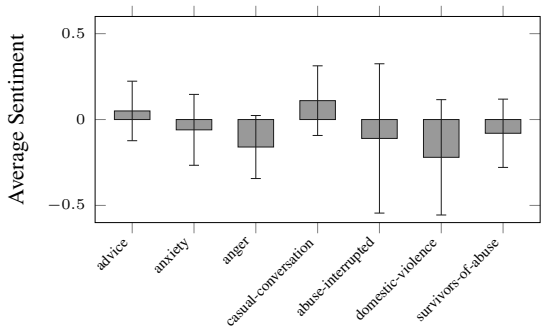

Figure 2: Mean VADER sentiment scores for posts within each subreddit. Error bars display the standard deviation of sentiment score.

Bansal (2016)) as a strategy to better understand the classification decision and feature discovery of the best-performing neural model. Activation clustering (Girshick et al., 2014) treats the activation values of n neurons per input as coordinates in a n-dimensional space. K-means clustering is then performed on them to group together inputs that maximally activate similar neurons. Each cluster can have a different pattern or trait in common.

One of the abuse-labeled clusters involves asking questions: *"how long after your abuse occurred did you have sex again?"*, *"what to do when you're falling behind: the other side of 'fear of missing out'."*, *"want to reduce mental illness? address trauma."*. In contrast, we also found a non-abuse cluster which was also a collection of question-based posts, but containing very different types of questions (e.g., about casual conversation topics, anxiety coping, holidays, family, etc.): *"What's your favorite Christmas song?"*, *"What are your favorite coping skills?"*, and *"How would your family and friends react to your reddit profile?"*. This difference demonstrates that the model is not only identifying grammatical patterns such as questions, but also learning some useful distinguishing linguistic characteristics and content of the abuse versus non-abuse stories. Moreover, the aggregation of these abuse-positive questions allows for a greater understanding of the needs of domestic abuse survivors. For example, the fields of psychology and therapy can potentially gain a more data-intensive and holistic view of what victims of domestic abuse have questions about and where to find answers.

6.2.1 Automatic Cluster Pattern Analysis

We also perform automatic pattern discovery inside different activation clusters, building on the manual analysis performed by Aubakirova and Bansal (2016). The most common words in each cluster were tallied. For example, an abuse-labeled cluster about questions has *"and"*, *"have"*, *"what"*, *"they"*, *"help"*, *"is"*, *"domestic"*, *"children"*, *"violence"*, and *"beaten"* in the top ten most common words. However, a non-abuse cluster about questions has *"do"*, *"what"*, *"your"*, *"and"*, *"I"*, *"it"*, *"would"*, *"have"*, *"to"*, and *"of"* as its top ten most used words. This shows that abuse-heavy posts contain more specific help and violence related words, whereas the non-abuse posts are very diverse and of several generic, casual topics. Because the top ten words in the abuse versus non-abuse clusters are different, we can also use these patterns to help classify future text rapidly. New phrases that exhibit similar common words to an already distinct cluster will more likely fall into that cluster of abuse or non-abuse.

7 Conclusion

This work applied three neural models—CNN, LSTM-RNN, and CNN-LSTM—to an important classification task in today's world of online social activism. We contributed a new state-of-the-art accuracy in this task and also presented interpretability techniques to understand neural feature discovery. Furthermore, we compared the sentiment scores of each subreddit and found that sentiment alone could not be used to classify stories as abuse-positive or abuse-negative.

This task can be used in the future as an aggregation tool to allow for the collection and linguistic explanation of meaningful stories from across the internet and social media platforms. The courage of the individuals who share personal domestic abuse stories must not be wasted. Instead of allowing this public outcry to remain imprisoned on the internet, let us use natural language processing to automatically amass and explain their stories, thus giving insights to helping victims of domestic abuse and allowing activists to spread awareness and enact real social change.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. This work was supported by a Google Faculty Research Award, a Bloomberg Data Science Research Grant, an IBM Faculty Award, and NVidia GPU awards.

References

Malika Aubakirova and Mohit Bansal. 2016. Interpreting neural networks to improve politeness comprehension. In *EMNLP*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

CJ Hutto Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Nicolas Schrading, Cecilia Ovesdotter Alm, Ray Ptucha, and Christopher Homan. 2015. An analysis of domestic abuse discourse on reddit. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2577–2583.

Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis Lau. 2015. A c-lstm neural network for text classification. *arXiv preprint arXiv:1511.08630*.