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

UNC DEPARTMENT OF COMPUTER SCIENCE

Sweta Karlekar and Mohit Bansal

UNIVERSITY OF NORTH CAROLINA CHAPEL HILL

Abstract

The detection and classification of domestic abuse stories shared online have 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**.

Task
Use NLP techniques to **classify and visualize/analyze/interpret** the linguistic characteristics of domestic abuse stories shared online.

- Develop **CNN, LSTM-RNN, and CNN-LSTM** neural models to detect domestic abuse stories in the Reddit Domestic Abuse dataset
- Achieve 95.8% accuracy in classifying whether posts contain abuse stories, **outperforming the current state-of-the-art accuracy**
- Assess feasibility of classification using only **sentiment scores**
- Present interpretable and explainable analysis of the neural model's predictions using **activation clustering techniques** to automatically discover linguistic features

Dataset

Schrading et al. (2015) assembled the Reddit Domestic Abuse dataset, which is publicly available. Reddit is a social media platform that contains a substantially large range of forums called subreddits in which users post comments pertaining to specific topics.

The Reddit Domestic Abuse dataset contains submissions of two types of labels:

- Abuse:
  - "*abuse-interrupted*"
  - "*domestic-violence*"
  - "*survivors-of-abuse*"
- Non-Abuse:
  - "*anger*"
  - "*anxiety*"
  - "*advice*"
  - "*casual-conversation*"

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

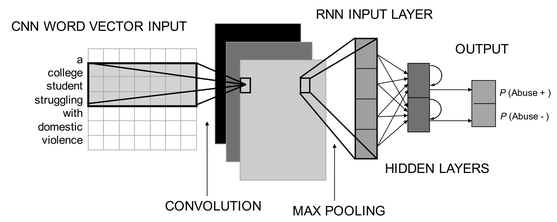
To balance the negative sentiment of abuse stories, submissions from "*anger*" and "*anxiety*" subreddits are included as non-abuse. Submissions from the "*advice*" subreddit are included as non-abuse to ensure that classifiers are not just finding help-seeking behavior. The subreddit "*casual-conversation*" is included as non-abuse as well. The dataset contains 1,336 total cases of abuse and 17,020 total cases of non-abuse.

Models

CNN: An embedding and a convolutional layer is applied, followed by a max-pooling layer. The convolution features are obtained by applying filters of varying window sizes to each window of words. The result is then passed to a softmax layer that outputs probabilities over two classes (i.e. abuse and non-abuse).

LSTM-RNN: Our LSTM-RNN model consists of an embedding layer followed by an LSTM layer. The final state, containing information from the entire sentence, is fed to a fully-connected layer followed by a softmax layer to obtain the output probabilities.

CNN-LSTM: Laying an LSTM-RNN layer on top of a CNN allowed us to combine both models' complementary strengths.



Results

The table below shows classification accuracies of related work (Schrading et al., 2015) 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%.

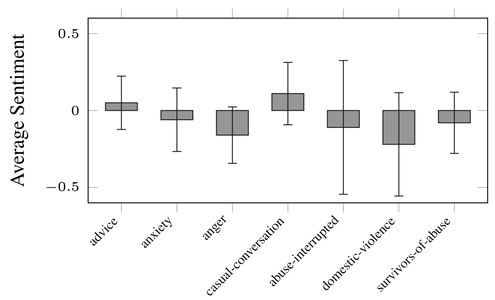
Model	Accuracy
Schrading et al. (2015)	92.0%
2D-CNN	92.6%
LSTM-RNN	94.5%
CNN-LSTM	**95.8%**

Analysis

Sentiment-Based Classification:

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.

- All of the abuse-positive subreddits had negative mean sentiment scores.
- Both the "anger" and "anxiety" subreddits (that were part of the non-abuse dataset) had negative average sentiment scores as well.
- Sentiment of each subreddit had large standard deviations. Overall, this demonstrates that **sentiment alone cannot be used to effectively classify stories** as abuse-positive or abuse-negative.



Activation Clustering Analysis:

Activation clustering treats the activation values of n neurons per input as coordinates in an n-dimensional space (Girshick et al., 2014; Aubakirova and Bansal, 2016). K-means clustering is then performed to group together inputs that maximally activate similar neurons.

We found abuse-positive and abuse-negative clusters that formed around questions:

Abuse Question Cluster:

- *"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."*

Non-Abuse Question Cluster:

- *"What's your favorite Christmas song?"*
- *"What are your favorite coping skills?"*
- *"How would your family and friends react to your reddit profile?"*

The aggregation of these abuse-positive questions allows for a better 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 the questions and concerns of victims of domestic abuse**.

Novel Automatic Pattern Discovery: The most common words in each cluster were tallied.

Abuse cluster questions have {"*and*", "*have*", "*what*", "*they*", "*help*", "*is*", "*domestic*", "*children*", "*violence*", "*beaten*"} as the top ten most common words.

Non-abuse cluster questions have {"*do*", "*what*", "*your*", "*and*", "*I*", "*it*", "*would*", "*have*", "*to*", "*of*"} as the top ten most common 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.

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 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 in a timely manner.

References

- Malika Aubakirova and Mohit Bansal. 2016. Interpreting neural networks to improve politeness comprehension. In *EMNLP*.
- C. Hutto and E. 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)*.
- 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.
- 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.

Acknowledgements

This work was supported by a Google Faculty Research Award, a Bloomberg Data Science Research Grant, an IBM Faculty Award, and NVidia GPU awards.