EMORY UNIVERSITY SCHOOL OF MEDICINE | Injury Prevention Research Center at Emory

Publication Brief | January 2023

Natural Language Model for Automatic Identification of Intimate Partner Violence Reports from Twitter

Intimate partner violence (IPV) is a preventable public health problem affecting millions worldwide. Approximately one in four women are estimated to be or have been survivors of severe violence at some point in their lives, irrespective of age, ethnicity, and economic status.
Survivors often report IPV experiences on social media, and automatic detection of such reports via machine learning may enable improved surveillance and targeted distribution of support and/or interventions for those in need. However, no artificial intelligence systems for automatic detection currently exist, therefore this study attempted to address this research gap.

Citation:
Al-Garadi, M. A., **Kim, S.**, Guo, Y., Warren, E., Yang, Y.-C., Lakamana, S., & **Sarker, A.**. (2021). *Natural Language Model for Automatic Identification of Intimate Partner Violence Reports from Twitter*. https://doi.org/10.1016/j.array.2022.100217

Correspondence:
Abeed Sarker
abeed.sarker@emory.edu

Methods:

- **Data Collection:** Researchers collected publicly available English posts (tweets) related to IPV from Twitter using its public streaming application programming interface.
- **Annotation guidelines:** Four annotators encoded each tweet into one of two categories —personal (reported by victims themselves), IPV-report (or IPV), or non-IPV-report (or non-IPV).
- **Text classification model:** Researchers investigated three different approaches for constructing an IPV classifier: traditional machine learning models, deep learning models, and transformer-based models. The primary objective was to create a model for identifying IPV tweets from streaming Twitter data.
- **Post classification analyses**: The analyses included a learning curve analysis, error analyses, and an analysis of biases.

Findings:

- The total number of annotated tweets was 6,348, with a considerable imbalance in the distribution (non-IPV tweets: 5,680 [~89%]; IPV tweets: 668 [~11%]).
- We divided the instances via stratified sampling across training, validation, and test sets by 70%, 10%, and 20%, respectively (training 4,443, validation 635, and test 1,270).
- The average pairwise IAA among double-annotated tweets (N = 1,834) was k = 0.86 (Cohen's kappa [28]), which can be interpreted as a substantial agreement.
- Our best performing transformer-based model achieved classification F1-scores of 0.76 for the IPV-report class and 0.97 for the non-IPV-report class.

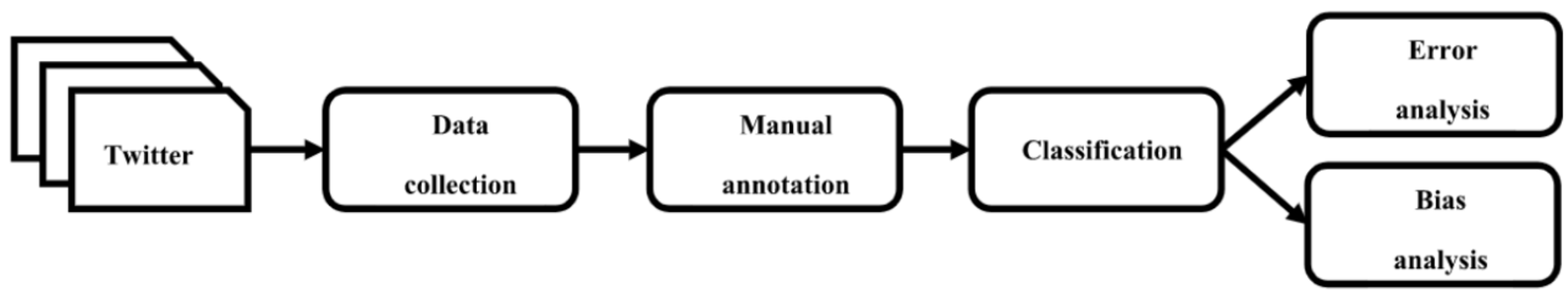

Fig.1. Shows the general framework describing overall process for developing NLP pipeline for Classifying IPV-related tweets

Discussion:

Although IPV survivors often reach out for support and information through social media channels such as Twitter, there is little effort to use such platforms to meet their needs. Posts about IPV are typically lost in the massive volume of data constantly posted on social media. Developing an effective, low-biased, and trustworthy model for classifying self-reported IPV in social media has significant practical applications. This study developed and evaluated an NLP pipeline to collect and classify posts from Twitter. Our NLP pipeline achieved comparable performance to humans and was particularly found to not have any biased on gender or race-related words. By identifying IPV survivors on Twitter, our model will lay the groundwork to design and deliver evidence-based interventions, support IPV survivors, and prevent and respond to IPV at scale and a low cost.