Array 15 (2022) 100217

Contents lists available at ScienceDirect

Array

journal homepage: www.sciencedirect.com/journal/array

Natural language model for automatic identification of Intimate Partner Violence reports from Twitter

Mohammed Ali Al-Garadi [a,*], Sangmi Kim [b], Yuting Guo [a], Elise Warren [c], Yuan-Chi Yang [a], Sahithi Lakamana [a], Abeed Sarker [a,d]

[a] *Department of Biomedical Informatics, School of Medicine, Emory University, Atlanta, GA, United States*
[b] *School of Nursing, Emory University, Atlanta, GA, United States*
[c] *Rollins School of Public Health, Emory University, Atlanta, GA, United States*
[d] *Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, United States*

ARTICLE INFO

Keywords:
Intimate partner violence
Domestic violence
Natural language processing
Machine learning
Social media

ABSTRACT

Intimate partner violence (IPV) is a preventable public health problem that affects millions of people worldwide. Approximately one in four women are estimated to be or have been victims of severe violence at some point in their lives, irrespective of age, ethnicity, and economic status. Victims 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 exists, and we attempted to address this research gap. We collected posts from Twitter using a list of IPV-related keywords, manually reviewed subsets of retrieved posts, and prepared annotation guidelines to categorize tweets into IPV-report or non-IPV-report. We annotated 6,348 tweets in total, with the inter-annotator agreement (IAA) of 0.86 (Cohen's kappa) among 1,834 double-annotated tweets. The class distribution in the annotated dataset was highly imbalanced, with only 668 posts (~11%) labeled as IPV-report. We then developed an effective natural language processing model to identify IPV-reporting tweets automatically. The developed model achieved classification F_1-scores of 0.76 for the IPV-report class and 0.97 for the non-IPV-report class. We conducted post-classification analyses to determine the causes of system errors and to ensure that the system did not exhibit biases in its decision making, particularly with respect to race and gender. Our automatic model can be an essential component for a proactive social media-based intervention and support framework, while also aiding population-level surveillance and large-scale cohort studies.

1. Introduction

Intimate partner violence (IPV) is a preventable public health problem that impacts millions of people worldwide [1]. IPV can be defined as physical or sexual assault, or both, of a spouse, partner, or cohabiting dating couples [2,3]. Approximately one in four women in the United States (US) is estimated to be or have been victims of severe violence at some point in their lives, irrespective of their age, ethnicity, and economic status [1]. IPV victims suffer from physical or mental health problems in the short and long term, including injury, pain, sleep problems, depression, post-traumatic stress disorder (PTSD), and suicide [4]. Family members and children who live in or witness violence can also experience adverse health and social developmental issues [5]. Notably, children exposed to IPV are more likely to experience depression, anxiety, PTSD, dissociation, and anger [6,7]. Physical health outcomes include injuries, sexually transmitted diseases, back and limb problems, memory loss, dizziness, and gastrointestinal conditions. Female victims could experience undesirable reproductive outcomes, including miscarriages and gynecologic disorders. From an economic perspective, the approximate IPV lifetime cost is $103,767 per female victim and $23,414 per male victim; a population economic burden amounts to nearly $3.6 trillion (2014 US$) over victims' lifetimes, 37% of which, $1.3 trillion, is paid by the government [8].

Emerging data show that since the outbreak of COVID-19, reports of IPV have increased worldwide [9]. In the US, IPV reports in the early phases of COVID-19 increased in many states and cities compared to prior years [10,11]. The current crisis, described as a "horrifying global surge in domestic violence" by the United Nations Chief [12], may be

[*] Corresponding author.
E-mail address: mohammed.a.al-garadi@vumc.org (M.A. Al-Garadi).

https://doi.org/10.1016/j.array.2022.100217
Received 11 January 2022; Received in revised form 1 July 2022; Accepted 2 July 2022
Available online 20 July 2022

attributed to mandatory lockdowns or movement restrictions to curb the spread of COVID-19 [12,13]; isolation of IPV victims with their abusers at home; elevated tension between partners, compounded by security, health, and financial worries (e.g., socioeconomic instability and business closures) [12,14]. The Sustainable Development Goals—a collection of 17 interlinked global goals designed to be a "blueprint for achieving a better and more sustainable future for all—focus on eliminating all forms of violence against women and girls [15,16].

Conventionally, IPV-related data have been collected through surveys and medical/police reports [17,18]. It has become challenging to obtain IPV-related data from these traditional sources during the COVID-19 pandemic because IPV victims, for example, were reluctant to see healthcare providers due to the fear of contracting the virus [17,18]. Social media websites (SMWs), such as Twitter and Reddit, can potentially get around the pandemic-induced obstacles by collecting data online. More than 4.5 billion individuals use SMWs globally, many of whom use it for extended periods [19]. SMWs have become a new communication tool for people to express their thoughts, emotions, opinions, and discuss daily problems, regardless of geographical locality. During the COVID-19 pandemic and lockdown, SMWs have become many people's primary channels to express and share information, emotions, and details of daily lives with their friends and family members [20]. SMWs may have particular utility to IPV victims because they tend to share their sensitive information more with friends (64%) and family members (49%), but less with healthcare providers (26%), police (23%), and shelter advocates (20%) [21,22]. Also, SMWs have been shown in past research to be excellent sources for collecting live data precisely, at scale (i.e., a large number of observations), anonymously, unobtrusively, and at a low cost [23,24]. Thus, SMW data about IPV and other digital footprints of IPV victims' behaviors can be analyzed in real-time to model patterns of IPV and discover underlying risk factors at the population and individual levels [25]. Moreover, during the COVID-19 pandemic, the United Nations urged to increase investment in online technology and civil society organizations to build harmless means for women to find help and support without informing their abusers, and accordingly reduce domestic violence [12]. SMWs, where users can share anonymous postings, may serve as secure platforms compared to other means (e.g., phone call, email) for offering instrumental, informational, and social support to IPV victims. These advantages of SMWs may complement (not substitute) the conventional resources (e.g., hotline, shelter) and strengthen the effort to prevent IPV and support IPV victims.

To the best of our knowledge, there has been no study employing social media analytics (e.g., natural language processing (NLP), machine learning) to identify IPV victims on SMWs for surveillance, prevention, and/or intervention. Therefore, this study aims to develop an automated, social media-based system to detect and categorize streaming IPV-related big data (into IPV and non-IPV cases) during the COVID-19 pandemic on Twitter via NLP and machine learning.

1.1. Significance

An effective NLP pipeline, including an automatic machine learning classifier, will help develop a surveillance system for IPV self-reported on SMWs. Such a pipeline may be used to detect and proactively reach out to large numbers of IPV victims simultaneously to provide support instead of waiting for them to seek help. Also, the pipeline will lay a foundation to potentially deliver evidence-based, non-contact interventions to IPV victims on SMWs for IPV prevention, mental health treatment, empowerment, and support [26].

1.2. Contributions

- We are the first study to collect data, develop annotation guidelines, and then manually annotate a large number of tweets related to IPV.
- We are also the first to develop an effective NLP pipeline involving supervised machine learning that can automatically extract and classify IPV-related tweets.
- We present a thorough analysis of the classification errors, and the model performances at different training data sizes, as this information may be crucial for future research.
- We present an analysis of potential biases and the trustworthiness of our model's performance through content analysis and the approach of layered integrated gradients.
- We describe challenges faced during our analysis and lessons learned, which will help future researchers design similar NLP systems for social media-based data.

2. Methods

2.1. Data collection

We collected publicly available English posts (tweets) related to IPV from Twitter using the SMW's public streaming application programming interface (API). The keywords we used are provided in Supplementary Table S1. We also used the Python library snscrape to collect data during the COVID-19 pandemic between January 1, 2020 – and March 31, 2021.

2.2. 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*). Such categorization of each tweet was made based on the definition of IPV. The Centers for Disease Control and Prevention defines IPV as physical violence, sexual violence, stalking, and psychological aggression (including multiple coercive tactics) by a current or former intimate partner (i.e., spouse, boyfriend or girlfriend, dating partner, or ongoing sexual partner) [27]. Supplementary Table S2 provides comprehensive details and examples of various types of IPV. Two necessary factors that determined if a tweet was a self-report of IPV were the (1) mention of an intimate partner as an abuser and (2) mention or description of any type of abuse (physical violence, sexual violence, stalking, and psychological aggression) or abusive tactics.

We conducted the annotation process iteratively over small datasets (~200 tweets). Guided by the definition of IPV and a domain-expert (SK), the annotators discussed disagreements in the early annotations until consistent coding rules were reached. With the finalized annotation guidelines, the annotators encoded the final dataset ($n = 6{,}348$) used for training and testing our NLP model. The gold-standard dataset was developed once reliable levels of agreement between the annotators were achieved. A subset of the tweets was annotated twice or thrice for computing inter-annotator agreements (IAA). For double-annotated tweets, if the annotators disagreed with the class, the tweet was assessed by an independent annotator who resolved the disagreement. The final IAA was calculated using Cohen's kappa [28].

2.3. Text classification model

We investigated three different approaches for constructing an IPV classifier. We used three *traditional* machine learning algorithms: decision tree [29], support vector machine (SVM) [30,31], and neural networks (NN) [32]. We also experimented with more advanced algorithms for text classification, including a deep learning-based algorithm, namely, bi-directional long short-term memory (BiLSTM) [33,34] and two transformer-based models, namely, Bidirectional Encoder Representations from Transformers (BERT) [35,36] and Robustly Optimized BERT (RoBERTa) [36].

Pre-processing stage: Before feeding the tweets to the model, they were cleaned by removing unwanted words, such as non-English characters. We replaced and anonymized the URLs and usernames with predefined tags (i.e., URL and <user>). We also lowercased all text.

Traditional machine learning models: For the traditional classifiers, we produced 1,000 most frequent *n*-grams (adjacent series of words with *n* ranging from 1 to 3: unigrams ($n = 1$), bigrams ($n = 2$), and trigrams ($n = 3$)) [37]. The stemmed and lowercased tokens from the texts were used to generated the n-gram vectors.

Deep learning model: We converted each word into a corresponding word vector then fed it into the BiLSTM classifier. For a word to vector conversion, we used the Twitter GloVe word embeddings trained on 2 billion tweets and 27 billion tokens, including a 1.2 million-word vocabulary. We used uncased GloVe embedding with 200-dimension vectors [38].

Transformer-based models: We used BERT and RoBERTa. The preprocessed training tweets were fed into the model for fine-tuning the model for IPV classification. The hyper-parameters and technical details are presented in Supplementary Table S3.

Model training validation: Our primary objective was to create a model for identifying IPV tweets from streaming Twitter data. Our primary metric for assessing classifiers' performance was the F_1-score (harmonic mean of precision and recall) for the IPV class. We divided the annotated dataset into training (70%), development (10%), and test sets (20%). We used the training dataset for training models, the development dataset for optimizing model hyperparameters, and the test set to evaluate and compare classifier performances.

2.4. Post-classification analyses

Learning curve analysis: We evaluated the model performance at different sizes of training data (20%, 40%, 60%, 80%, 100%) to investigate the behavior of each model at a given training percentage and assess whether growing the training dataset improved the models' performance on the fixed test dataset.

Error analysis: To identify potential reasons for misclassifications, we studied the common patterns of errors made by our model by manually analyzing the contents of a subset of misclassified tweets.

Bias analysis: Generally, humans recognize the meanings of sentences by focusing on the most important words. Some words play more important roles than others in understanding a post's meaning and deciding its class (e.g., IPV or non-IPV). We aimed to examine the words on which our best performing model focuses and which ones have more importance than the others in making the classification decisions. This process helps us ensure that the developed model has a low bias, is trustworthy, and understands the model's errors. To achieve this objective, we used the approach of layered integrated gradients [39] proposed by Captum [40,41]. The approach is based on an attribution technique called integrated gradients [42], which uses an interpretable algorithm that calculates word importance scores by approximating the integral of the gradients of the model's outputs with respect to the inputs [42,43]. We analyzed a random sample of 5% of the posts from the test set in this manner.

3. Results

3.1. Annotation and inter-annotator agreement

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 [44] (see Fig. 1).

3.2. Classification results

The classifier performances on the test set are shown in Table 1. Accuracies and class-specific F_1 scores are reported for each classifier. The confusion matrix for the best-performing classifier (i.e., RoBERTa) is shown in Fig. 2. As Table 1 illustrates, the BERT model also obtained competitive results. However, traditional machine learning and deep learning models did not perform well, particularly for the primary evaluation metric (IPV-report F_1 score), with no significant differences between these classifiers.

3.3. Learning curve

Fig. 3 shows the learning curve at different percentages of training data (20%, 40%, 60%, 80%, 100%) and the performance obtained using fixed test data. Overall, the performance tends to improve with increasing training data, which is expected, particularly from 20% to 80%. With only 40% of the training dataset, the pre-trained models (i.e., BERT, RoBERTa) performed similarly to the traditional and deep learning models with 100% of the training dataset.

3.4. Error and model analyses

Given that the RoBERTa model (Table 1) produced the best performance compared with other models, we used it for our NLP pipeline and investigated the errors made by this classifier. For the tweets annotated as *non-IPV-report*, few non-IPV examples were classified as IPV as reflected by a very high F_1-score for the non-IPV class (i.e., 0.97). Nevertheless, in a few cases, an error occurred when tweets contained users' indirect IPV experiences (reporting IPV not experienced personally but by others). For instance, "*my ex-neighbours rowed a lot. he was a big bloke, over day i heard him hit her; my heart beating out my chest, I knocked on their door. it stopped that fight. < hashtag> domestic violence.*" Additionally, non-IPV examples were misclassified as IPV when tweets contained hypothetical scenarios not experienced by the authors in the real world, such as: "*that is completely different issue. people not wanting to have children is different from divorce. i could have children with good intentions in mind, and still get divorced bc my husband turns abusive.*" In contrast, for the tweets annotated as IPV tweets, misclassifications often occurred into non-IPV tweets when the author's IPV was implied in the tweet but was not explicit. For instance, "*I hate how abusers justify their abuse of their partner in their head. That was what my ex would do. the physical scars heal, but the mental ones remain.*" In this example, the author did not state up front that her ex-partner harmed her body and mind. "*my ex would do*" did not contain any indicator of IPV; hence, the tweet was inaccurately considered to be non-IPV by our model.

3.5. Model behavior and bias analysis

We first checked whether our model's classification outcomes were biased toward a specific gender or racial group. We made this our focus because many recent studies have shown that machine learning models

Table 1
Accuracies and F_1 scores on the test set for the traditional classifiers (decision tree, SVM, NN) and advanced transformer-based classifiers (BERT, RoBERTa).

Classifier	Accuracy (%)	IPV F_1 score	Non-IPV F_1 score
Decision tree	89	0.48	0.94
SVM	93	0.52	0.96
Neural network	91	0.50	0.95
BiLSTM	91	0.44	0.95
Transformer (BERT)	94	0.70	0.97
Transformer (RoBERTa)	**95**	**0.76**	**0.97**

SVM = Support Vector Machine, NN = Neural Network, BiLSTM = Bi-Directional Long Short-Term Memory, BERT = Bidirectional Encoder Representations from Transformers, RoBERTa = Robustly Optimized BERT.

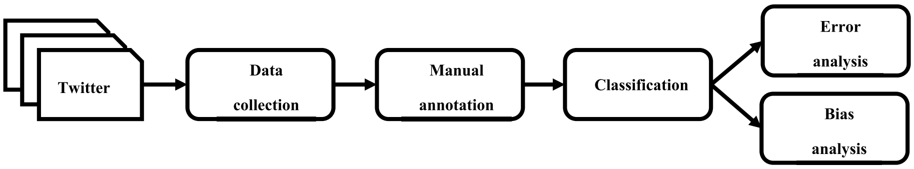


Fig. 1. Shows the general framework describing overall process for developing NLP pipeline for classifying IPV-related tweets.

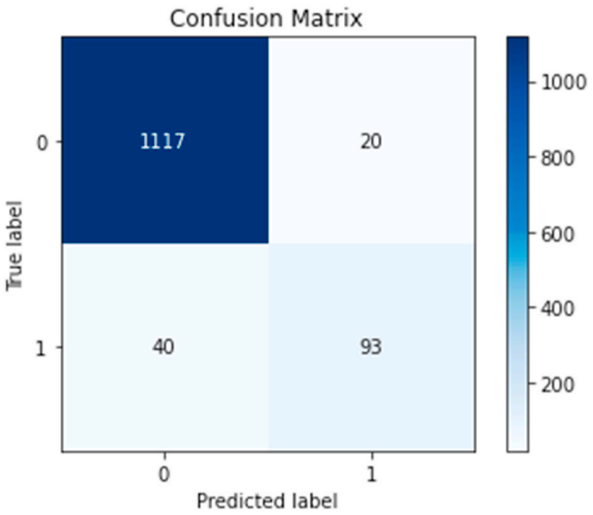


Fig. 2. Confusion matrix for the best classifier (RoBERTa) on the IPV binary classification task.

often utilize race or gender information when making inferences rather than more relevant data [45,46]. We selected IPV tweets that our model correctly classified and also contained words indicating the authors' gender identities (e.g., man, woman). We checked whether the model focused on male- or female-related words to make the classification decisions. We used a similar approach to IPV tweets that our model misclassified. Simultaneously, we changed male-related words into female-related words and vice versa to examine if the change in the gender-based word would also alter the misclassification outcome or if this change significantly affected the word importance of the gender-based word in the tweet. We applied the same method to race-related words (e.g., White, Black). The results showed that our model's classification decisions were not biased by gender or race. Table 2 presents examples of this analysis.

We also analyzed the misclassified tweets further in terms of word importance to understand the model's focus during the classification process. We found that the model often assigned importance to words that were not useful to human annotators (an example is provided in Table 2). For instance, in the example in Table 2, the model focuses on the words " < *user* > , < *repeat* > ," which means mentioning other users in the post and repeating the previous word, were considered to be important by the model. These features are not important to human annotators for this task. Importantly, however, the model did not make misclassifications based on demographic information such as race and gender.

3.6. Use of the model on a large dataset

We were able to retrieve 153,826 tweets using IPV-specific keywords, as described in the previous section. We observed that most social media users used terms like "domestic violence" and "IPV" when discussing the topic. The hashtag "#domesticviolence" was also frequently used. Given that the main objective of the study is to identify self-reports of IPV victims, we further narrowed our pipeline by considering only those tweets which mentioned personal pronouns ("I," "my," "me," "us"). Finally, 39,393 tweets from 29,583 users were collected and analyzed for content.

With the application of classification models, a total of 1,803 reports were found, comprising approximately 4.6% of the collected tweets. Further text analysis on the positive tweets showed the types of abuses that users reported and how they handled them. *N*-gram distributions helped identify the frequently discussed topics. Stop words and other commonly used terms related to this study, such as "domestic violence," "retweet," "behavior," "victim," and "violence," were removed to reduce noise, followed by lemmatization. The results of the analysis showed that most of the users described how they were victimized, and the top repeating mentions were abusive relationships (9.23%), threatening or attempting to kill (3.5%), sexual assault (2.77%), and child abuse (2.21%) (Fig. 4). In contrast, as shown in Fig. 5, some users expressed

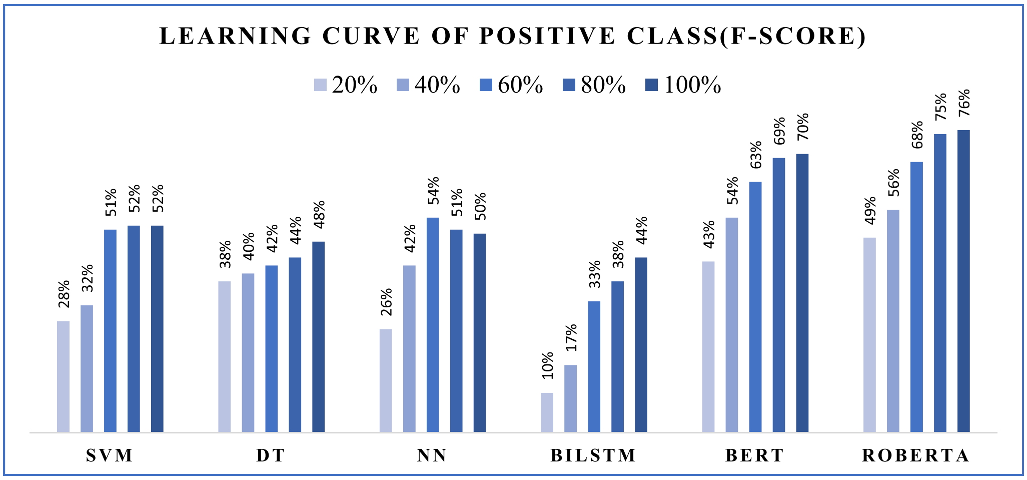
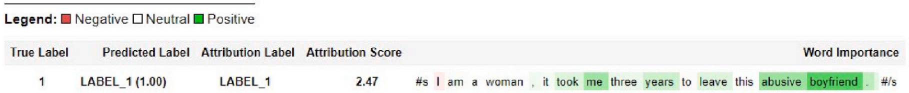
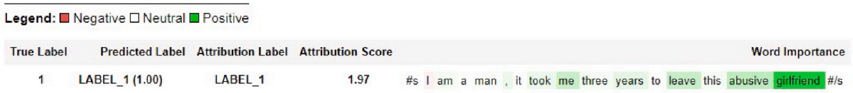
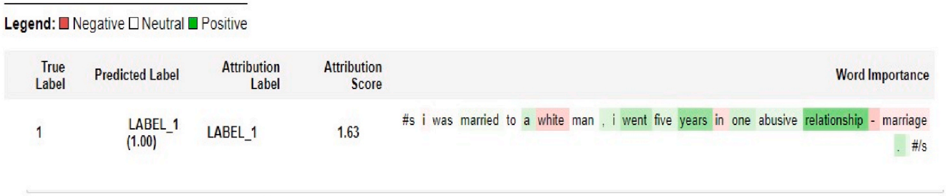
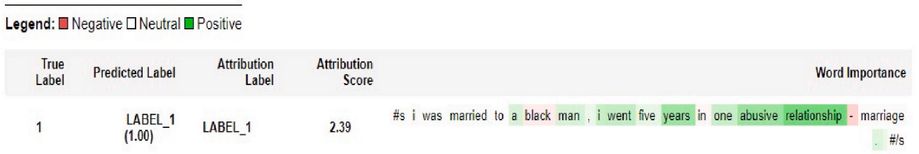
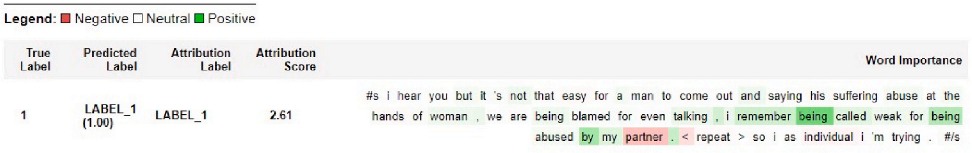
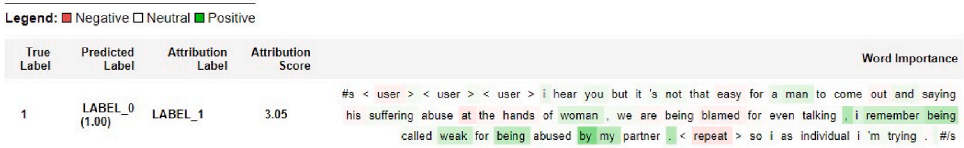


Fig. 3. Classifier performances at different training set sizes (learning curves). SVM = Support Vector Machine, DT = Decision tree, NN = Neural Network, BiLSTM = Bi-Directional Long Short-Term Memory, BERT = Bidirectional Encoder Representations from Transformers, RoBERTa = Robustly Optimized BERT.

Table 2
Bias and model outcome analyses were conducted using the layered integrated gradients approach. Highlighted text segments show where the model focused when making classification decisions. IPV (Positive Class), Non-IPV(Negative class).

Example	Human annotation	Model prediction	Classification results and word importance	Comment
Gender bias analysis				
I am a woman, it took me three years to leave this abusive boyfriend.	Positive (1)	Positive (1)	Legend: 🟥 Negative □ Neutral 🟩 Positive True Label Predicted Label Attribution Label Attribution Score Word Importance 1 LABEL_1 (1.00) LABEL_1 2.47 #s I am a woman , it took me three years to leave this abusive boyfriend . #/s Legend: 🟥 Negative □ Neutral 🟩 Positive True Label Predicted Label Attribution Label Attribution Score Word Importance 1 LABEL_1 (1.00) LABEL_1 1.97 #s I am a man , it took me three years to leave this abusive girlfriend #/s	Label correctly classified, and the classification outcomes did not change when we changed the gender-related word in the example. The main important word for classifying positive class (green highlights) is still the same in both examples.
Race bias analysis *" i was married to a white man, i went five years in one abusive relationship-marriage."*	Positive (1)	Positive (1)	Legend: 🟥 Negative □ Neutral 🟩 Positive True Label Predicted Label Attribution Label Attribution Score Word Importance 1 LABEL_1 (1.00) LABEL_1 1.63 #s i was married to a white man , i went five years in one abusive relationship - marriage . #/s Legend: 🟥 Negative □ Neutral 🟩 Positive True Label Predicted Label Attribution Label Attribution Score Word Importance 1 LABEL_1 (1.00) LABEL_1 2.39 #s i was married to a black man , i went five years in one abusive relationship - marriage . #/s	Label correctly classified, and the classification outcomes did not change when we changed the race-related word in the example. The main important word for classifying positive class (green highlights) is still the same in both examples.
Importance of irrelevant words analysis *⟨user⟩ ⟨user⟩ ⟨user⟩ i hear you but it's not that easy for a man to come out and saying his suffering abuse at the hands of woman, we are being blamed for even talking, i remember being called weak for being abused by my partner. <repeat⟩so i as individual i'm trying.*	Positive (1)	Negative (0)	Legend: 🟥 Negative □ Neutral 🟩 Positive True Label Predicted Label Attribution Label Attribution Score Word Importance 1 LABEL_1 (1.00) LABEL_1 2.61 #s i hear you but it 's not that easy for a man to come out and saying his suffering abuse at the hands of woman , we are being blamed for even talking , i remember being called weak for being abused by my partner . < repeat > so i as individual i 'm trying . #/s Legend: 🟥 Negative □ Neutral 🟩 Positive True Label Predicted Label Attribution Label Attribution Score Word Importance 1 LABEL_0 (1.00) LABEL_1 3.05 #s < user > < user > < user > i hear you but it 's not that easy for a man to come out and saying his suffering abuse at the hands of woman , we are being blamed for even talking , i remember being called weak for being abused by my partner . < repeat > so i as individual i 'm trying . #/s	The model is still sensitive to irrelevant social media features or language. Removing word ⟨user⟩ change the classification outcome from negative (incorrect class) to positive class (correct class).

their personal experiences of handling violent acts, such as restraining orders (5.5%) and seeking help from police (3.3%). A comparison of male and female self-reports showed that the number of tweets mentioning ex-husbands is three times as many as those about ex-wives, although the distributions of their mentions are similar in the full data set.

4. Discussion

The RoBERTa classifier achieved the best performance compared to other models, and the performance was almost comparable to human annotators (based on the IAA). This suggests that the developed systems can be practically employed for collecting and analyzing IPV data at a larger population level. Findings from the learning curve analysis illustrate the usefulness of the pretrained model (e.g., RoBERTa) in learning from a small dataset. The pretrained model continues improving in performance with additional training data but with a slow learning rate between 80% and 100%. It is likely that further manual annotation of data will improve performance, but the rate of improvement for the IPV-report class will be slow. The classification outcomes and word importance analyses show that our system is not biased toward certain gender or race groups. However, this result does not necessarily mean that our developed model is bias-free, which is difficult to claim without a large test data that contains more information. Rather, we intended to ensure that the model at this stage was not making classifications based on gender- or race-related words, as well as to find the best strategies to improve the quality and diversity of training datasets in the future.

4.1. Limitations and future directions

Based on our analysis and observations, we identified several limitations and future directions:

Our model tends to misclassify tweets that contain implicit self-reports of IPV. Future work may include strategies to enrich the classification with new training datasets that address this issue. Enriching the training data can be planned to mainly boost the number of tweets with similar characteristics to those misclassified in the first training round. The classifier also often misclassifies tweets due to words and symbols that uniquely appear in social media data, such as "<," and <user>, as well as the tweets that contain subtle language cues that are not captured in the training data. Annotating a larger data set for training is also likely to resolve some of the performance issues related to these. It must be noted, however, that manual annotation is an extremely time consuming process and this acts as a limitation to intelligent system development.

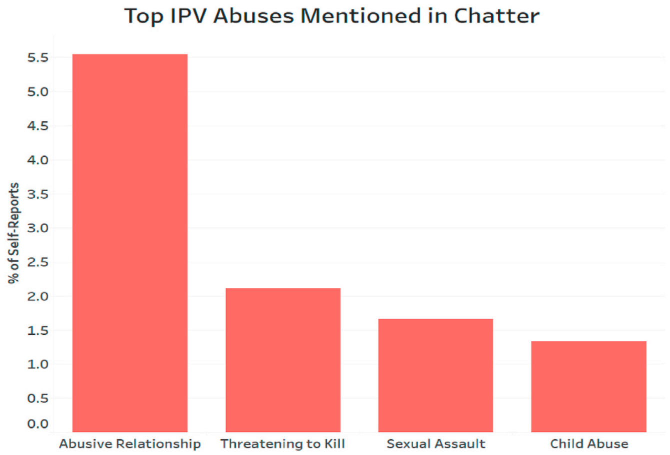


Fig. 4. of self-reports for different types of IPV abuses (top four cases of abuse of IPV).

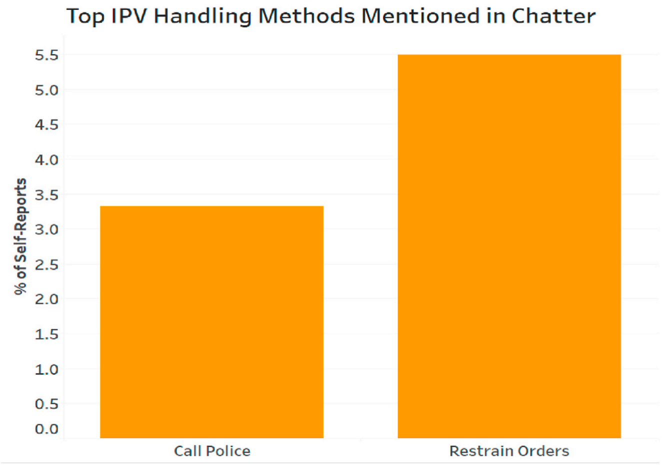


Fig. 5. Percentage of self-reports about different ways of IPV handlings mentioned in Twitter.

Some word importance scores by our model are not correlated with the word importance estimates of human annotators. This may at least partially be related to our data collection strategy. We used words such as "domestic violence" or "abuse" to collect our data. Therefore, positive (IPV-report) and negative (non-IPV-report) examples in our training and test data have similar distributions of these words. Consequently, the model does not give these words high importance. Furthermore, although such strategies had been essential to extract relevant tweets only, we are likely to miss the IPV tweets that do not contain any of our IPV-related words. To improve our approach in this regard, we will attempt to incorporate more expansive strategies for data collection. Recent works on topics similar to IPV may serve as the best initial sources to explore. While there are currently no datasets, other than ours, specific to IPV from social media, some recent studies have explored topics such as sexism [47], misogyny [48], and anti-women hate speech [49,50]. In some cases, IPV can be viewed as a violent case of sexism or misogyny, and so leveraging the strategies described in these related research works may help us expand our IPV dataset.

Even though our model does not show any observable bias toward gender and race while making its classification outcomes, further examination is required before and during the deployment stage to keep the model unbiased and trustworthy and prevent any data drift that may cause undetectable biases. It is specifically important to analyze posts that our model correctly classifies to ensure that the correct classification decisions are not driven by biased reasoning. Content analyses of large sets of posts also need to be carried out routinely to identify potential problems with the classification approach. In this paper, we performed a brief content analysis, which helped us discover important information such as types of IPV. Importantly, it did not reveal any observable biases associated with the data that we have gathered. Note, however, that because of the keyword/phrase-based data collection strategy, there is always the possibility of introducing data-specific biases. It can be particularly difficult to ascertain in such approaches to data curation what important contents may be missing from the data. In the future, we intend to execute conduct analyses to discover data-specific biases introduced due to the use of the keyword-based collection, similar to the one described in Poletto et al. (2021) [50].

Although not identified through content analysis, we suspect the insufficient contexts in the description of the events as possible reasons for misclassifications. Twitter as a micro-blogging platform allows only a limited number of characters in one tweet. Therefore, users often attempt to embed all their experiences in short sentences, forgoing detailed contexts that would be useful for accurate classification into either IPV-report or not. Users often overcome this limitation by posting a series of tweets (thread) explaining their IPV (or other) experiences.

The current strategy of automatic classification does not take into account neighboring posts in threads. In the future, we will explore strategies for incorporating thread-level information in the classification process.

The dataset used to pre-train the RoBERTa or the BERT model mainly contains a corpus of books (800 million words) and English Wikipedia. The way people express themselves in SMWs is different from the texts in books [37]. Efforts have been made to pre-train BERT-related models on social media. However, previous research showed that the RoBERTa model based on traditional text still performs better than these models in several social media NLP tasks [51,52]. A potential direction may be combining traditional data with social media data to have a hybrid model built with a large text corpus and social media language.

5. Conclusion

Although IPV victims often reach out for support and intervention through social media channels such as Twitter, there is little effort to use such platforms to address this public health problem. 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 the Twitter platform to identify IPV-related tweets. Our NLP pipeline achieved comparable performance to humans and was shown to be particularly not biased to gender or race-related words. By identifying IPV victims on Twitter, our model will lay the groundwork to design and deliver evidence-based interventions, and support to IPV victims and potentially enable us to address the problem of IPV in close to real-time via social media.

Authors' contributions

MAA, **SK** and **AS** designed the experiments **MAA**, **SK**, **YCY**, **YG**, **EW**, **SL** and **AS** conducted the data collection, analysis, and evaluations. **MAA**, **SK** and **AS** conducted manuscript writing. **MAA**, **SK**, EW, and AS helped interpret relevant findings, discussed and analyzed the results.

Funding statement

The study was funded by the Injury Prevention Research Center at Emory (IPRCE), Emory University.

Declaration of competing interest

The authors have declared no competing interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.array.2022.100217.

References

[1] Smith SG, et al. The national Intimate Partner and Sexual Violence Survey (NISVS): 2015 Data Brief – Updated Release. Atlanta, GA: National Center for Injury Prevention and Control, Centers for Disease Control and Prevention; 2018.

[2] Campbell JC. Health consequences of intimate partner violence. The lancet 2002; 359(9314):1331–6.

[3] Capaldi DM, Knoble NB, Shortt JW, Kim HK. A systematic review of risk factors for intimate partner violence. Partner abuse 2012;3(2):231–80.

[4] G. Dillon, R. Hussain, D. Loxton, and S. Rahman, "Mental and physical health and intimate partner violence against women: a review of the literature," International journal of family medicine, vol. 2013, 2013.

[5] Karakurt G, Koç E, Çetinsaya EE, Ayluçtarhan Z, Bolen S. Meta-analysis and systematic review for the treatment of perpetrators of intimate partner violence. Neurosci Biobehav Rev 2019;105:220–30.

[6] Cummings EM, Davies PT. Maternal depression and child development. JCPP (J Child Psychol Psychiatry) 1994;35(1):73–122 [Online]. Available: https://acamh.onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-7610.1994.tb01133.x?sid=nlm%3Apubmed.

[7] Johnsona RM, et al. Adverse behavioral and emotional outcomes from child abuse and witnessed violence. Child Maltreat 2002;7(3):179–86.

[8] Peterson C, et al. Lifetime economic burden of intimate partner violence among U. S. Adults. Am J Prev Med 2018;55(4):433–44. https://doi.org/10.1016/j.amepre.2018.04.049 (in eng).

[9] Kim S, Sarker A, Sales JM. The use of social media to prevent and reduce intimate partner violence during COVID-19 and beyond. Partner Abuse 2021;12(4):512–8.

[10] Police Bureau Portland. Trends Analysis: Pre & Post School Closures – April 15, 2020. 2020. Accessed: August 20, 2020. [Online]. Available: https://www.portlandoregon.gov/police/article/760237.

[11] Cuomo Governor Andrew M. Following spike in domestic violence during COVID-19 pandemic, secretary to the governor melissa derosa & NYS council on women & girls launch task force to find innovative solutions to crisis. accessed August 20, 2020), https://www.governor.ny.gov/news/following-spike-domestic-violence-during-covid-19-pandemic-secretary-governor-melissa-derosa.

[12] United Nations. In: UN chief calls for domestic violence 'ceasefire' amid 'horrifying global surge'. UN News; 2020.

[13] Boserup B, McKenney M, Elkbuli A. Alarming trends in US domestic violence during the COVID-19 pandemic. Am J Emerg Med Apr 28 2020. https://doi.org/10.1016/j.ajem.2020.04.077 (in eng).

[14] B. Gosangi et al., "Exacerbation of physical intimate partner violence during COVID-19 lockdown," Radiology, vol. 0, no. 0, p. 202866, doi: 10.1148/radiol.2020202866.

[15] Agüero JM. COVID-19 and the rise of intimate partner violence, vol. 137. World development; 2021, 105217.

[16] T. E. Union, "Ending Violence Against Women and Girls," The Spotlight Initiative, Accessed 11/16/2021. [Online]. Available: https://www.un.org/sustainabledevelopment/ending-violence-against-women-and-girls/.

[17] Gosangi B, et al. Exacerbation of physical intimate partner violence during COVID-19 pandemic. Radiology 2021;298(1):E38–45.

[18] Moreira DN, Pinto da Costa M. The impact of the Covid-19 pandemic in the precipitation of intimate partner violence (in eng) Int J Law Psychiatr 2020;71. https://doi.org/10.1016/j.ijlp.2020.101606. 101606-101606, July-August.

[19] Kepios, Global Social Media Statistics. *online source Accessed June 26 2022,* 2022. [Online]. Available: https://datareportal.com/social-media-users#:~:text=Analysis%20from%20Kepios%20shows%20that,since%20this%20time%20last%20year.

[20] Koeze E, Popper N. The virus changed the way we internet. In: The New York times; 2020.

[21] Glass N, Eden KB, Bloom T, Perrin N. Computerized aid improves safety decision process for survivors of intimate partner violence. J Interpers Violence 2010;25 (11):1947–64. https://doi.org/10.1177/0886260509354508.

[22] Glass N, Eden KB, Bloom T, Perrin N. Computerized aid improves safety decision process for survivors of intimate partner violence, vol. 25; 2010. p. 1947–64. no. 11.

[23] Schwab-Reese LM, Hovdestad W, Tonmyr L, Fluke J. The potential use of social media and other internet-related data and communications for child maltreatment surveillance and epidemiological research: scoping review and recommendations (in eng) Child Abuse Negl Nov 2018;85:187–201. https://doi.org/10.1016/j.chiabu.2018.01.014.

[24] Lin H, et al. User-level psychological stress detection from social media using deep neural network. Orlando, Florida, USA. In: Presented at the Proceedings of the 22nd ACM international conference on Multimedia; 2014. https://doi.org/10.1145/2647868.2654945 [Online]. Available:.

[25] Merchant RM, Lurie N. Social media and emergency preparedness in response to novel coronavirus. JAMA 2020;323(20):2011–2. https://doi.org/10.1001/jama.2020.4469.

[26] El Morr C, Layal M. Effectiveness of ICT-based intimate partner violence interventions: a systematic review. BMC Publ Health 2020;20(1):1–25.

[27] Breiding M, Basile KC, Smith SG, Black MC, Mahendra RR. Intimate partner violence surveillance: Uniform definitions and recommended data elements. CDC report; 2015., Version 2.0.

[28] Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull Oct 1968;70(4):213–20. https://doi.org/10.1037/h0026256.

[29] Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics 1991;21(3):660–74.

[30] Chang CC, Lin CJ. LIBSVM: a library for support vector machines. Acm T Intel Syst Tec 2011;2(3):1–27. https://doi.org/10.1145/1961189.1961199 (in English) Artn 27.

[31] Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers 1999;10(3): 61–74.

[32] Wang S-C. Artificial neural network. In: Interdisciplinary computing in java programming. Springer; 2003. p. 81–100.

[33] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8). https://doi.org/10.1162/neco.1997.9.8.1735. 1735-80, Nov 15.

[34] Schuster M, Paliwal KK. Bidirectional recurrent neural networks. IEEE Trans Signal Process 1997;45(11):2673–81.

[35] Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018. *arXiv preprint arXiv:1810.04805*.

[36] Liu Y, et al. Roberta: A robustly optimized bert pretraining approach. 2019. *arXiv preprint arXiv:1907.11692*.

[37] Al-Garadi MA, et al. Text classification models for the automatic detection of nonmedical prescription medication use from social media. BMC Med Inf Decis Making 2021;21(1):1–13.

[38] Pennington J, Socher R, Manning CD. Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing. EMNLP; 2014. p. 1532–43.

[39] Mudrakarta PK, Taly A, Sundararajan M, Dhamdhere K. Did the model understand the question?. 2018. *arXiv preprint arXiv:1805.05492*.

[40] Kokhlikyan N, et al. Captum: A unified and generic model interpretability library for pytorch. 2020. *arXiv preprint arXiv:2009.07896*.

[41] Pierse C. Transformers Interpret. 2021 [Online]. Available: "https://github.com/cdpierse/transformers-interpret.

[42] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: International conference on machine learning. PMLR; 2017. p. 3319–28.

[43] Hayati SA, Kang D, Ungar L. Does BERT Learn as Humans Perceive? Understanding Linguistic Styles through Lexica. 2021. *arXiv preprint arXiv:2109.02738*.

[44] Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med May 2005;37(5):360–3 (in eng).

[45] Lwowski B, Rios A. The risk of racial bias while tracking influenza-related content on social media using machine learning. J Am Med Inf Assoc 2021;28(4):839–49.

[46] DeBrusk C. The risk of machine-learning bias (and how to prevent it). MIT Sloan Management Review; 2018.

[47] Abburi H, Parikh P, Chhaya N, Varma V. Semi-supervised multi-task learning for multi-label fine-grained sexism classification. In: Proceedings of the 28th international conference on computational linguistics; 2020. p. 5810–20.

[48] Anzovino M, Fersini E, Rosso P. Automatic identification and classification of misogynistic language on twitter. In: International conference on applications of natural language to information systems. Springer; 2018. p. 57–64.

[49] Frenda S, Ghanem B, Montes-y-Gómez M, Rosso P. Online hate speech against women: automatic identification of misogyny and sexism on twitter. J Intell Fuzzy Syst 2019;36(5):4743–52.

[50] Poletto F, Basile V, Sanguinetti M, Bosco C, Patti V. Resources and benchmark corpora for hate speech detection: a systematic review. Comput Humanit 2021;55(2):477–523.

[51] Guo Y, Dong X, Al-Garadi MA, Sarker A, Paris C, Aliod DM. Benchmarking of transformer-based pre-trained models on social media text classification datasets. In: Proceedings of the the 18th annual workshop of the australasian language technology association; 2020. p. 86–91.

[52] Guo Y, Ge Y, Al-Garadi MA, Sarker A. Pre-trained transformer-based classification and span detection models for social media health applications. In: Proceedings of the sixth social media mining for health (# SMM4H) workshop and shared task; 2021. p. 52–7.