HHS Public Access

Author manuscript

J Comput Soc Sci. Author manuscript; available in PMC 2025 April 29.

Published in final edited form as:

J Comput Soc Sci. 2022 November ; 5(2): 1207–1233. doi:10.1007/s42001-022-00166-8.

Identification of intimate partner violence from free text descriptions in social media

Phan Trinh Ha[1], **Rhea D'Silva**[2], **Ethan Chen**[2], **Mehmet Koyutürk**[1,3], **Günnur Karakurt**[4]

[1]Department of Computer and Data Sciences, Case Western Reserve University, Cleveland, OH, USA

[2]Department of Psychology, Case Western Reserve University, Cleveland, OH, USA

[3]Center for Proteomics and Bioinformatics, Case Western Reserve University, Cleveland, OH, USA

[4]Department of Psychiatry, Case Western Reserve University, Cleveland, OH, USA

Abstract

Intimate partner violence (IPV) is a significant public health problem that adversely affects the well-being of victims. IPV is often under-reported and non-physical forms of violence may not be recognized as IPV, even by victims. With the increasing popularity of social media and due to the anonymity provided by some of these platforms, people feel comfortable sharing descriptions of their relationship problems in social media. The content generated in these platforms can be useful in identifying IPV and characterizing the prevalence, causes, consequences, and correlates of IPV in broad populations. However, these descriptions are in the form of free text and no corpus of labeled data is available to perform large-scale computational and statistical analyses. Here, we use data from established questionnaires that are used to collect self-report data on IPV to train machine learning models to predict IPV from free text. Using Universal Sentence Encoder (USE) along with multiple machine learning algorithms (random forest, SVM, logistic regression, Naïve Bayes), we develop DetectIPV, a tool for detecting IPV in free text. Using DetectIPV, we comprehensively characterize the predictability of different types of violence (physical abuse, emotional abuse, sexual abuse) from free text. Our results show that a general model that is trained using examples of all violence types can identify IPV from free text with area under the ROC curve (AUROC) 89%. We also train type-specific models and observe that physical abuse can be identified with greatest accuracy (AUROC 98%), while sexual abuse can be identified with high precision but relatively low recall. While our results indicate that the prediction of emotional abuse is the most challenging, DetectIPV can identify emotional abuse with AUROC above 80%. These results establish DetectIPV as a tool that can be used to reliably detect IPV in the context of various applications, ranging from flagging social media posts to detecting IPV

Phan Trinh Ha, pxt177@case.edu, Mehmet Koyutürk, mxk331@case.edu, Günnur Karakurt, gkk6@case.edu.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42001-022-00166-8.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

in large text corpuses for research purposes. DetectIPV is available as a web service at https://www.ipvlab.case.edu/ipvdetect/.

Keywords

Data analysis; Public health; Psychology; Data visualization; Social media

Introduction

Intimate partner violence (IPV) is a public health problem that affects millions of individuals worldwide. IPV is considered as any behavior by a current or former partner that causes harm to those involved in the relationship [1]. IPV includes physical, sexual, and emotional forms of violence and aggressive acts [2]. Physical violence is defined as the use of physical force and actions to inflict physical harm in a partner. These actions include, but are not limited to, slapping, pushing, kicking, punching, cutting, burning, and using weapons. Sexual violence can be defined as forcing or attempting to force a partner to take part in a sex act, sexual touching or a non-physical sexual event when the partner does not or cannot consent. Emotional violence is assessed by measuring harms to emotional well being.

According to the World Health Organization, it is estimated that on average 30% of women experience IPV globally [3]. Based on the National Intimate Partner and Sexual Violence Survey, about 1 in 4 women as well as 1 in 10 men reported experiencing some form of IPV in the US [2, 4]. Unfortunately, IPV is also among the leading causes of death among younger women [1, 5]. Statistics indicate that about one third of all women murdered in the United States are killed by their partner [6].

IPV causes numerous physical, psychological, sexual, and emotional trauma. Past research showed that victims of IPV are more likely to suffer from physical, mental, and sexual health issues in acute or chronic manner, than non-victims [7, 8]. Health problems ranging from minor cuts to broken bones, injuries, disability and other severe health consequences due to IPV are frequently reported by the victims [9]. Furthermore, mental health issues such as post-traumatic stress disorder, clinical depression, and suicide attempts are highly prevalent among the victims of IPV [7, 10, 3]. Recent research also indicates that emotional abuse caused by degradation, intimidation, and control, is highly prevalent in intimate relationships and co-occur frequently with physical violence [11]. For these reasons, effective identification of IPV in various settings can be useful in raising awareness, providing help to victims, and generating data for research to further understand the causes, consequences, and treatment of IPV.

Detection and measurement of IPV.

Detecting IPV in various settings can be challenging. Over time, various measurement approaches have been used by the scientific community to assess the nature and extend of IPV in relationships. These approaches heavily relied on measuring frequencies and occurrence rates of specific violent behaviors. The most widely used scale to assess both the perpetration and victimization is the Conflict Tactics Scale (CTS) [12, 13]. CTS measures

frequencies of perpetration and victimization of physical, emotional, and sexual violence that happened in the last year. Other scales such as the Abuse Behavior Inventory (ABI) [14] use variations of the CTS as behavioral checklists that define abuse acts through specific behaviors.

Emotional abuse is a complex construct that includes various emotional tactics by the abuser for continuous emotional mistreatment of a person. It can be defined as the use of verbal and non-verbal communication with the intent to harm another person mentally or emotionally and/or to exert control over another person. Acts demonstrating emotional abuse are thought to be along a spectrum, with deliberate scaring, humiliation, and isolation as well as threats to an individual's safety. Due to the complexity and the spectrum of emotional abuse, as well as its nonverbal components, researchers, and clinicians face challenges in consistently assessing emotional abuse. Widely used emotional abuse scales such as the Psychological Maltreatment Inventory (PMI) [15] and the Women's Experience with Battering (WEB) scale [16] define emotional abuse also through specific behaviors.

Measuring sexual violence in intimate relationships also faces many challenges as the definition is often viewed to be synonymous with rape. However, sexual violence consists of a spectrum of behaviors such as sexual demands that make partners uncomfortable, pressuring, controlling, and manipulating to obtain sex and degradation of partners. To understand the extent of the problem in intimate relationships, scales such as Sexual Experiences Survey (SES) ask questions regarding sexual experiences and assign participants into categories of non-victimized, sexually coerced, abused, or assaulted based on the severity of sexual violence [17].

Use of social media to characterize public health problems.

With the growing popularity of social media platforms, the amount data generated from user posts and activities along with the availability of tools to collect user information is more than ever before. As a result, patterns and indicators of public health problems can be extracted from the social media, in an attempt to better characterize these problems, assess their prevalence, and identify their association with other issues.

In recent literature, there have been many applications of using social media to characterize public health problems from large amounts of data generated from social media platforms [18]. Social media data have been successfully utilized in public health concerns including investigating adverse drug reactions [19], infectious/virus monitoring [20], and in understanding the effects of smoking cessation and substance use interventions [21]. In relation to problems of substance abuse, social media platforms such as Instagram and Twitter provide a multitude of data formats to characterize, in the form of both text and images posted by users on social media platforms [22]. For abusive behaviors, there have also been usages of data from political sites [23] and YouTube comments [24], the prior noted to have collected over 10 million data points for analysis [23]. Social media data have also been utilized to detect the use of hate speech and even radical violence [25]. Using Yahoo! data and Part-of-Speech tagging to perform hate speech classification [26], a small feature set performed better in detecting hate speech as compared to the global set. Previous researchers also used topic modeling as well for identification of hate speech [27].

Images can be used in conjunction with image tagging tools to generate description tags for user's posts and can then be used for inferring a local region's public health problem. [28]. With the abundance of free floating text in users post, text data are also commonly used in conjunction with domain expertise for encoding to characterize behaviors [18, 22].

Challenges in predicting IPV from social media.

In identifying public health problems, sets of keywords can be used as identifiers. However, IPV can become more context-dependent and could be described by victims or perpetrators in different ways. In this case, use of keywords alone can overlook cases of IPV, and as such should require a use of a higher level representation to describe. Thus more sophisticated machine learning and natural language processing (NLP) algorithms can be potentially more effective in detecting IPV in social media data. However, there are no existing labelled datasets with regards to IPV to train sophisticated machine learning algorithms. In particular, there is no readily available corpus of accounts of victims of IPV cases. As a result, manual curation and labelling is required to generate datasets for training and testing machine learning models to predict IPV in social media data.

Our solution and contributions.

To overcome the challenges posed by data, identification, and model building, we present a comprehensive framework with the following components:

- We use established questionnaires from the literature as training data. This provides a corpus for training machine learning models in addition to the readily trained models that we build and validate in this study.
- We use natural language embedding based text embedding models to enable training of standard representation learning algorithms on free text. Importantly, the embeddings can be computed based on larger corpuses that are not necessarily labeled.
- We use Reddit's `r/relationship_advice` platform as our data source from social media for testing and validation. For this purpose, we use a list of keywords curated by our domain experts to generate a set of candidate posts from `r/relationship_advice`, and use manual labeling to characterize the existence/description of violence in each sentence per post. We also randomly select posts/sentences from other subreddits to generate different types of negatively labeled samples. Besides serving as test data for the validation of our machine learning models, these labeled data also provide a valuable resource for data-driven discovery using other computational approaches.
- We use a variety of machine learning algorithms, collectively named DetectIPV, to comprehensively characterize the predictability of violence. Using this framework, we develop general as well as type-specific violence models to characterize different types of IPV.
- To make DetectIPV available to the public, clinicians, and the scientific community, we release DetectIPV as a web service with an easy-to-use interface

and visualizations that allow exploratory analyses. The web service is available at http://www.ipvlab.case.edu/ipvdetect/.

Methods

Study design

The purpose of our study is to identify intimate partner violence (IPV) from free text. In our computational experiments, we focus on social media posts, but the text can also be in the form of scientific articles, free-floating notes from clinicians, and case interviews. While the posts can be (and usually are) composed of multiple sentences, we consider sentences as units of analysis as this simplifies the task of learning.

Labeling of sentences.

IPV can include multiple types of violence including physical, emotional, and sexual abuse. To capture the specificity of types of violence, we categorize sentences using five different labels:

- **General violence (GV):** The sentence describes a situation that involves IPV.
- **Physical abuse (PA):** The sentence describes a situation that involves physical abuse. PA is defined as any form of intentional force with the potential for causing death, disability, injury or harm. It includes behaviors when a person hurts or tries to hurt a partner by hitting, kicking, or using another type of physical force. It can also be acts of physical assault, use of weapons and threats of assault by a partner.
- **Emotional abuse (EA):** The sentence describes a situation that involves emotional abuse. EA involves the continual emotional mistreatment of a person. It is the use of verbal and non-verbal communication with the intent to harm another person mentally or emotionally and/or to exert control over another person. Emotional abuse can involve deliberately trying to scare, humiliate, isolate, or ignore a person. It refers to behaviors that harm a person's self-worth or emotional well-being. Emotional abuse is the form of abuse that is not physical, rather emotional. This can include the diminishing of self-worth or identity, lack of independence, neglect, and extends to verbal abuse.
- **Sexual abuse (SA):** The sentence describes a situation that involves sexual abuse. Sexual abuse/violence is defined as any unwanted sexual activity, making unwanted sexual advances or abusing lack (or incapability) of consent. It is any sexual act, attempt to obtain a sexual act or other act directly against a person's sexuality using coercion, by any person regardless of their relationship to the victim in any setting. Sexual violence not only includes abuse/assault or rape but also includes sexual exploitation, sexual harassment, stalking, and cyber harassment. For the purpose of this research, we focus on experiences between two people in a relationship.
- **Negative (NE):** The sentence does not describe a situation that involves IPV. We consider two types of sentences that are considered as negative: (i) Unrelated

negatives are sentences that are irrelevant to relationships or violent situations. (ii) Related negatives describe situations that involve relationship problems but do not include violence or abuse, or describe healthy and/or happy relational interactions (Table 1)

Physical abuse, emotional abuse, and sexual abuse are subcategories of general violence that are not mutually exclusive. In other words, if a sentence is labeled GV, then it needs to be also labeled by at least one of PA, EA, or SA. If a sentence is labeled NE, then it cannot be labeled with any of PA, EA, or SA.

Machine learning task.

We set up the machine learning task as follows: Given a sentence, decide whether it describes a citation that involves IPV, in the form of one or more of physical, emotional, or sexual violence. In our framework, we train "General Violence" models to distinguish the category GV from the category NE, as well as "Type-Specific" models to distinguish each of PA, EA, and SA from NE. We use the distinction between unrelated and related negatives to characterize what the models are learning (e.g., are they learning how to distinguish relationship problems from irrelevant situations or are they learning to distinguish violent situations in relationships from non-violent situations in relationships?)

Computational workflow.

The workflow of our framework is shown in Fig. 1. For training, we obtain items from self-report questionnaires that are used in the literature to identify, quantify, or characterize violence in relationships. We then use Universal Sentence Encoder (USE) [29] to generate vector-space embeddings of these questionnaire items. Subsequently, we use the embeddings and labels of the questionnaire items to train predictive models In parallel, we perform literature review to identify keywords that are associated with IPV ("Terms of Interest"). We use these keywords to identify social media posts that are potentially associated with IPV ("Top Posts"). We manually investigate these posts to label each sentence in the post as positive or negative, also annotating the positive sentences with the type of violence. These sentences form our test data. We generate vector-space embeddings for test sentences using Universal Sentence Encoder (USE) [29] as well. Using these embeddings and the machine learning models trained on questionnaire items, we predict the labels of the test sentences and evaluate the predictive accuracy of the models.

Dataset retrieval and curation

Training data: questionnaires.—For training, we utilize self-report questionnaires that are used in the literature to assess conflict and/or violence in a relationship. The items in these questionnaires represent descriptions of a wide range of violent interactions that are commonly accepted as examples of IPV by the scientific community. Use of questionnaires for training serves two purposes: (i) the training data are reliable (based on scientific literature), (ii) all the data obtained from social media can be for testing. This is important as social media data are not readily labeled, and thus requires a large amount of manual work for labeling. The questionnaires that are used to obtain training data are shown in Table 2.

Positively labeled samples for training data (GV, PA, EA, SA) are items from the CDC's compendium of assessment tools for IPV [78]. The items representing related negatives represent sentences containing words that also exist in IPV-associated sentences. For example, the sentence "`My partner told me that I am worthless`" describes emotional abuse, but the word "worthless" can be contained in the sentence "`I feel worthless`", which is not an indicator for IPV. The unrelated negatives are questionnaire items that are irrelevant to IPV and other relationship issues, taken from multiple unrelated topic questionnaires.

Overall, the training data include 355 positively labeled and 707 negatively labeled sentences. Among the positively labeled sentences, 126 are labeled as physical abuse, 198 are labeled as emotional abuse, and 31 are labeled as sexual abuse. Among the negatively labeled sentences, 410 are unrelated negatives while 297 are related negatives. All sentences that are used as positive and negative training samples are provided in Supp. Table 1.

Test data: social media posts.—The test data are collected from the subreddit `r/relationship_advice` from the social media site Reddit, from posts that were created between January 2019 and July 2020, using PushShift's API. After processing, This dataset contains 116 posts composed of a total of 326 sentences.

To identify posts that are potentially related to IPV, we use a word-frequency based approach. For this purpose, we manually curate a list of 119 terms that are potentially related to IPV. These keywords are listed in Supp. Table 2. Using these terms, we generated a set of candidate posts that can potentially describe IPV situations. We then manually evaluate these posts to identify sentences that describe situations involving IPV.

The process of generating the set of candidate posts is as follows:

- Lemmatize all words in posts and terms so that all words are simplified to their basic form (e.g., both `punching` and `punches` are reduced to punch).
- Compute the frequency of all 119 terms from all collected posts. Let $f(t)$ denote the frequency of term t in the corpus.
- Compute the score for each post D as $score(D) = \sum_{t \in D} \frac{1}{\log_2(f(t))}$. Use of inverse term frequency ensures diversification of the posts, i.e. posts with uncommon
- keywords are not overwhelmed by posts with more common keywords.
- Rank the posts in decreasing order of $score(D)$. For manual evaluation, we focus on the 90 posts with highest scores. We call these posts candidate posts.

Each candidate post is then evaluated and labeled by two experts, sentence by sentence, to match our machine learning framework. For this purpose, we develop a detailed manual for the operational definitions of different types of IPV, and their underlying concepts. These operationalizations with multiple examples helps the team identify the presence of different types of violence in qualitative descriptions of relational experiences.

Team members are trained in recognizing the signs of the relational problems over an hour long weekly meetings for 6 months. When there were disagreements among the independent reviewers, these sentences were brought to the larger team for discussion. To investigate discrepancies, operational definitions were reviewed for each sentence, emotional, physical, and sexual abuse to clarify and explore discrepancies. The disagreements were discussed until the larger team agreed on all interactions.

During the process of evaluation and labeling, our team found that sentences with IPV can contain more than one type of IPV. For example, the sentence "`No matter where we were, when she was angry she would not hesitate to verbally and physically abuse me`", contains both physical abuse and emotional abuse. Consequently, in our labeling, we allow a positively-labeled sentence to be assigned more than one type of IPV (PA, EA, SA). As a result, we obtain 326 positively-labeled test sentences, of which 74 are labeled as PA, 257 are labeled as EA, and 25 are labeled as SA (with some sentences containing overlaps).

We obtain "Unrelated negative" test sentences by extracting posts from an unrelated subreddit. Namely, we collect 196 posts on `r/changemyview` and select a random sentence from each post. For "Related negative" test sentences, we randomly selecte sentences from the original corpus we obtained from `r/relationship_advice`, which are verified by the curation team to ensure that these sentences are relationship related, but do not describe any IPV-related situations. As a result, we obtain 292 negatively labeled test sentences, of which 196 are considered unrelated negatives and 96 are considered related negatives. The list of positive and negative test sentences is provided in Supplemental Table 3.

The resulting number of training and test sentences and the distribution of their labels are shown in Table 3.

Model building

Computation of sentence embeddings.—As it is difficult to interpret unstructured data, transformation of unstructured data into a more structured format is usually desirable. This is often accomplished by embedding "unstructured" samples into a low-dimensional vector space [79, 80]. In the context of natural language processing, there are algorithms for embedding words, sentences, or documents. Since our focus here is on sentences, we use Universal Sentence Encoder (USE) [29], a sentence encoder with documented success in various applications [81–83]. Given a sentence, USE uses neural network architectures that are pre-trained on large corpuses of text to identify latent dimensions that are descriptive of the variance in these corpuses. This is particularly useful in our application as it removes the requirement for abundant data to compute robust and descriptive embeddings.

USE works on a per-sentence basis, and is designed specifically for transfer learning, where the model's output can be reused as part of another task. USE takes in a string sentence and output a fixed 512-dimensional vector representing the sentence. These embeddings take into account sentence semantics, such that sentences with similar meanings are closer to each other in the embedding space. Proximity in the embedding space is quantified in terms of the angular distance between the embeddings u and v of a pair of sentences:

$$\mathrm{sim}(\mathbf{u},\mathbf{v}) = \left(1 - \arccos\left(\frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}\right)/\pi\right). \tag{1}$$

USE has two different model architectures, one with Deep Averaging Network (DAN) [84] and the other with a Transformer [85]. These models are pre-trained on larger sets of data, thus they are ideal for our set-up as the training data that are available from questionnaires are relatively limited. In our application, we use a pre-trained DAN-based encoder. We chose to use the DAN-based encoder as it provides better runtime performance compared to the Transformer based encoder with minimal accuracy loss, which is helpful for the model to perform inference on the web-hosted DETECTIPV.

In this architecture, the output token embeddings, which are dense vectors that correspond to a token, are averaged. USE implements this architecture by generating embeddings from each word and bi-gram token, which are then passed through a deep feed-forward network to compute the sentence embedding as a non-linear combination of these word and bi-gram embeddings.

Feature selection.—The embeddings provided by USE do not take into account our training data and map the sentences to a fixed 512-dimensional space. For this reason, it is potentially useful to order these dimensions in terms of their discriminative potential based on the training data and perform feature selection. We take a filtering-based approach to feature selection and rank the sentences according to their confusion in distinguishing positive training examples from negative training examples. For this purpose, we formulate a penalty function that assesses each dimension's ability to separate positive examples from negative examples in the training data: Letting $\mathscr{D}$ denote the set of all dimensions returned by USE and $\mathscr{N}$ denote a given set of negative examples we define the *separability* of a pair p_i and p_j of positive samples with respect to a given set of dimensions $D \subseteq \mathscr{D}$ and a as:

$$separability^{(D)}(p_i, p_j \mid \mathscr{N}) = 1 - \frac{\sum_{n_k \in \mathscr{N}} I\left(sim^{(D)}(p_i, n_k) > sim^{(D)}(p_i, p_j)\right) + I\left(sim^{(D)}(p_j, n_k) > sim^{(D)}(p_i, p_j)\right),}{2 * \|\mathscr{N}\|}. \tag{2}$$

Here $sim^{(D)}(u, v)$ denotes the similarity between the embeddings of two samples in the space represented by D and I denotes indicator function. If D contains a single dimension, we use the negative absolute difference between u and v to compute $sim^{(D)}(u, v)$. If D contains more than one dimension, we use cosine similarity to compute $sim^{(D)}(u, v)$ (Equation (1)).

Then, for a given set of positive examples $\mathscr{P}$ and negative examples $\mathscr{N}$, we quantify the separability of $\mathscr{P}$ and $\mathscr{N}$ with respect to D as the mean of separability across all unique positive pairs $p_i, p_j \in \binom{\mathscr{P}}{2}$ from which can create $\binom{|\mathscr{P}|}{2}$ pairs

$$Separability^{(D)}(\mathscr{P} \mid \mathscr{N}) = \frac{\sum_{p_i, p_j \in \binom{\mathscr{P}}{2}} separability^{(D)}(p_i, p_j \mid \mathscr{N})}{\binom{|\mathscr{P}|}{2}}. \quad (3)$$

With these measures in hand, we compute the separability of all positive sentences and negative sentences in the training data with respect to each USE dimension $d \in \mathscr{D}$. We order the USE dimensions in decreasing order of $separability^{(\{d\})}(\mathscr{P} \mid \mathscr{N})$, and compute the $separability^{(D)}(\mathscr{P} \mid \mathscr{N})$ where D is a growing set of dimensions with each dimension added in order of individual separability. We then select the set of dimensions where separability on training data peaks and use those dimensions to train the models. Comprehensive results on feature selection and the effect of the number of embedding dimensions on predictive performance are shown in Sect. 3.4.

Training machine learning models.—To build a predictive model using USE embeddings of the sentences from questionnaires, we use four different machine learning algorithms. These algorithms are Logistic Regression, Naive Bayes, SVM and Random Forest. We then evaluate the ability of the model in detecting IPV in sentences extracted from Reddit posts. We pick these algorithms as they have been successfully applied in the past with similar applications, including mental statement assessment in social media [86]. We use social media data from Reddit for testing to investigate how well the models learned from the questionnaires generalize to practically relevant text data.

Type-specific vs. General violence models.—In addition to training a general model to identify intimate partner violence in general, we also train type-specific models to identify specific types of IPV, namely physical abuse (PA), emotional abuse (EA), and sexual abuse (SA). While the training for the general violence model incorporates all positively labeled samples, we train the type-specific models by considering the positively labeled samples that are labeled with the respective subtype. As negative training samples, both the general and type-specific models use the negatively-labeled samples. For each sentence, each trained model produces a confidence value indicating the likelihood of the sample belonging to a type of IPV. We use these confidence values to visualize the sentences in the space of the three types of IPV and explore the relationship between each subcategory of IPV.

Results

In all of our computational experiments, we use questionnaire data for training and social media data for testing. Our computational model uses sentences as units of analysis. The number of sentences in the training and testing datasets, and their distribution into violence subtypes and types of negatives are shown in Table 3.

Predictive performance of the general violence model

We first assess the predictive performance of the general violence model using four different classification algorithms. In this model, sentences with violence in the training

(questionnaire) data are labeled as “positive” with no distinction in terms of the type of violence and sentences with no violence are labeled as “negative” (including related and unrelated negatives). We also consider all positive and negative testing examples in testing, with no distinction with respect to type. The results of this analysis are shown in Fig. 2.

Overall, we observe that the models built using four different classification algorithms deliver similar predictive accuracy, where the area under ROC curve (AUROC) is slightly less than 90% for all algorithms. The waterfall plots provide further insights into the predictions of Logistic Regression and Random Forest Models, in which the positive testing examples are colored according to violence type. In these waterfall plots, the test sentences are ordered on the x-axis according to the confidence assigned by the classifier in labeling the sentence as “violence” and this confidence value is shown on the y-axis. As seen in these plots, both classifiers are able to assign high confidence scores to sentences with physical abuse, while confusion mostly occurs between sentences that contain emotional abuse and negatives. We also observe that the Logistic Regression Model is slightly more successful in distinguishing sexual abuse from negatives.

Predictive performance of type−specific violence models

As the results shown in the previous section demonstrate, the general violence model provides nearly 90% AUROC in detecting IPV in a sentence. When we consider the accuracy of this model in detecting specific types of violence, we observe that it is most successful in bringing forward physical abuse, while its performance is variable in detecting emotional or sexual abuse. This observation leads to the question of whether a type-specific model that is trained using only the specific violence type for positive examples can provide better performance in detecting the target violence type.

To answer this question, we compare the performance of the type-specific violence model in detecting their target violence type against that of the general violence model. The results of this analysis using Logistic Regression are shown in Fig. 3(a).

As seen in Fig. 3(a), the type-specific physical abuse model performs better than the general violence model in detecting physical abuse. This is also true for sexual abuse, where the type-specific model can detect sexual abuse with 90% AUROC. In contrast, the type-specific model for emotional abuse does not improve the detection of emotional abuse model as compared to the general model that also makes use of examples of physical and sexual abuse.

To investigate the reason behind the observed performance of type-specific violence models, we visualize the waterfall plots for each of the type-specific models using Logistic Regression and Random Forest. These results are shown in Fig. 3(b)–(e). We observe that the EA-specific model distributes confidence scores uniformly across all test sentences, the PA-specific model scores sentences with physical abuse with very high confidence, while assigning considerably low confidence scores to a majority of the test sentences (most of which are true negatives). The SA-specific model, on the other hand, tends to assign lower confidence scores to all test sentences—this is likely due to the relative scarcity of sentences with sexual abuse in our training and test data.

Visualization of sentences in the space of violence types

To gain further insights into what the algorithms learn to detect IPV in sentences, we visualize the predictions of the three type-specific violence models in the 3-dimensional space of violence types. For this purpose, we use the radial visualization technique developed by Hacıaliefendioğlu et al. [11]. The results of this analysis for Random Forest and Logistic Regression models are shown in Fig. 4. In the figure, the predictions of the models for training sentences are visualized on the left. For these sentences, the relative location of each point (representing a sentence) shows how much the model is able to fit the sentence to its true label (since these sentences are used in training). The web service implementation of DETECTIPV visualizes the query sentences in this radial space, providing users with the ability to visually explore the confidence of the models for each sentence.

As seen in Fig. 4, the Random Forest model is able to fit the data in a way that separates different types of samples in the space of violence types. Logistic Regression, on the other hand, incorporates the overlap between physical and emotional abuse, as well as the confusion between negative sentences and emotional abuse into the model. Interestingly, however, as shown on the right panel of the figures, the two models behave similarly on test sentences. Namely, the models can separate physical abuse from other types of violence (or negatives), while there is a lot of overlap between physical and emotional abuse predictions, negative sentences are mostly confused with emotional abuse, and the models assign low confidence to sexual abuse in general.

To investigate the underlying reasons for the difference in the behavior of Logistic Regression (LR) vs. Random Forest (RF) models, we further inspect the overall distribution of the confidence scores assigned to training and testing samples by each model. The results of this analysis are shown in Fig. 3(b–e). As seen in the figure, during training, LR fits a model that assigns confidence scores to training sentences across a dynamic range, whereas RF tends to fit a model that has a sharper decline for confidence scores assigned to negative training samples. On the testing data, this translates into many positive sentences being assigned lower scores by RF, while LR is able to use the dynamic range that is learned in training to distinguish the test samples. This may be due to the nature of the difference between the sentences used in training and testing. Namely, the sentences in the questionnaire (training) are more direct and shorter, whereas the sentences collected from social media (testing) can be longer and convoluted. For this reason, while the RF model is able to fit the questionnaire data very well, this model does not generalize to the social media data as well as the LR model.

Comparison against alternate approaches and effect of number of dimensions

To understand the benefits of using sentence embedding, we compare the predictive accuracy of DETECTIPV against a term-frequency based approach, in this case, TF-IDF with single words and bi-grams. While doing so, we also investigate the effect of the number of dimensions/features used for classification. For Universal Sentence Encoder, the features are the dimensions in the embedding space. To prioritize and select these features, we use a filtering approach using the separation of labeled training (questionnaire) sentences, as described in Sect. 2. For the term-frequency based approach, the features are words

and bigrams. To prioritize and select the term-frequency based features, we compute the mutual information between the presence of each term with sentence labels in the training (questionnaire) data and select features in decreasing order of mutual information. The results of this analysis are shown in Fig. 5.

As seen in Fig. 5, the predictive performance of DetectIPV goes up with increasing number of USE dimensions, and stabilizes at around 128 dimensions. This is also captured by our measure of separability on the training data, the separability of physical and emotional violence peaks at around 128 dimensions as well. Importantly, all four classification algorithms perform fairly similarly in detecting general violence when presented with the same set of features. It is interesting to note that these models perform fairly well (above 80% AUROC) even with the first 8 dimensions dimension, suggesting that the first 8 dimensions selected from USE provides a representative latent feature for the presence of violence in a sentence.

In contrast to DetectIPV, the TF-IDF models perform sligthly better than random (around 50% AUROC) when they utilize a low number of dimensions (terms), although the first few terms are most informative on violence as quantified by mutual information. As the number of terms go up, the predictive accuracy of TF-IDF models improves, and nears 75% AUROC when all words and bigrams are used, except for the Naïve Bayes model. Overall, these results show that the use of sentence embeddings in DetectIPV outperforms TF-IDF based models in detecting violence in sentences.

The effect of negative examples used in training

As discussed in Sect. 2, we use negative examples that are unrelated to relationships, as well as those that are on relationships, but do not include violence or abuse. Since DetectIPV potentially has a broad range of application domains, one can be interested in using it to identify violence in a corpus on relationship problems or identify IPV-related sentences in a larger, broader corpus. Motivated by this consideration, we comprehensively characterize the predictive accuracy of DetectIPV in distinguishing IPV-related sentences from sentences that are related or unrelated to relationships. The results of this analysis are shown in Table 4.

In Table 4(a), the results of cross-validation for 5 runs of 5-fold cross validation are shown. As seen on the table, all algorithms provide AUROC above 90% for all combinations of negative sentences used in training and testing. However, we observe that the AUROC goes down and variability in the performance of the Random Forest and Naïve Bayes models goes up when unrelated negatives are considered in training.

The generalizibility of the models, i.e., the AUROC achieved by models trained on questionnaire data and tested on social media data is shown in Table 4(b). As would be expected, DetectIPV delivers best predictive performance (above 90% AUROC for all models) in distinguishing IPV-related sentences from unrelated sentences when trained using unrelated negative examples. These models' AUROC goes down to nearly 80% when they are tested on a set that includes only relationship-related sentences as negatives. This observation suggests that models trained using unrelated negatives learn to detect presence

of violence, but they also learn to detect relationship issues to a certain extent. However, training on relationship-related sentences does not significantly improve the performance of this algorithms in distingushing IPV-related sentences from relationship-related sentences (AUROC around 80% for all models).

Importantly, in distinguishing IPV-related sentences from unrelated sentences, the four models that are trained using both related and unrelated negatives perform nearly as good as models that are trained only using unrelated negatives.

These results show that DetectIPV is quite robust to the types of sentences that are used in training and testing, and almost always delivers nearly 80% accuracy. These results also provide insights into selecting negative examples to train DetectIPV depending on the application..

Discussion

IPV has many devastating consequences on the victim's emotional and physical well being with substantial morbidity and mortality. IPV does not only involve tangible violence but also involve abusive situations that are harmful to emotional well being. These abusive situations can often be unrecognized and under-reported [87, 88] Our results show that the general model that is trained using examples of all violence types can identify IPV from free text with high (> 90%) confidence. We also train type-specific models and observe that physical abuse can be identified with high precision, while sexual abuse can be identified with high precision but relatively low recall. Our results also indicate that the prediction of emotional abuse is the most challenging.

Our model can distinguish physical abuse more distinctly as compared to other violence types. It is possible that signs of physical abuse are more noticeable. It is usually more clearly indicated as the language used to describe physical abuse is usually quite specific, such as hitting, scratching, pushing, shoving, throwing, grabbing, choking, biting, hair-pulling, slapping, hitting, punching, burning or use of restraints/body size and/or strength against another person. Furthermore, physical abuse is usually a patterned behavior in that it is usually not an isolated incident, but rather becomes gradually more frequent and also co-occurs with emotional abuse.

Our results indicate that emotional abuse is the hardest type of violence to distinguish among the free text. Emotional abuse, the continual emotional mistreatment of a person, diminishes a person's self-worth or emotional well-being and is associated with the development of depression, anxiety and post-traumatic stress disorder (PTSD) [89]. Emotional abuse is also considered a precursor to physical and sexual abuse. Identifying and interpreting emotional abuse presents a challenge due to the broad range of behaviors that may fall into this category and the need for contextualization.

Emotional abuse is often not visible, it can be passed and unnoticed in conversation or posts. It is often harder to identify indicators of emotional abuse online as the tone of the post is not as easily understood. Those who are not aware of their situation will change the language they use when describing their experience. This may be confusing

to our model and can trigger false positives or false negatives. However, certain words can imply a level of emotional abuse including but not limited to threat/threatened, control, fear/afraid, manipulate, neglect, demean, gaslit/gaslighting. In addition, the use of semantic embedding (Universal Sentence Encoder) can help capture context to a certain extent. Thus the sentence-focused approach of DetectIPV is able to detect emotional abuse with reasonable and potentially useful accuracy. However, approaches that use broader context (e.g., multiple sentences) and more sophisticated nature-language processing (NLP) algorithms can improve the accuracy of detecting emotional abuse in free text.

Our results indicate that when distinguishing sexual abuse, results indicate the importance of looking at a wider range of key terms, which may not be individually associated with violence. In general, sentences involving sexual abuse indicate both the physical and emotional nature of sexual abuse in intimate relationships.

The threat to violence and verbalizing these threats could be warning signs for severe and dangerous violence. Campbell's [30] danger assessment survey indicates that such threats often increase the risk of danger and homicide. Threats may also happen simultaneously with emotional, sexual, and physical abuse. It is possible that our model might incorrectly label threats as one or another form of violence. Similarly, although considered as emotional abuse, property damage can also involve physical violence as it can be viewed as symbolic violence and comfort with destructive power. [90] Relationships involving threats of violence and property damage can be labeled more accurately utilizing context-dependent information.

Limitations.

An important limitation of this study is the scale of data that is available for training and testing. Specifically, for violence types, training data were more scarse for Sexual abuse, due to the availability of a limited number of measures with good psychometric qualities. An important factor that contributes to the accuracy of machine learning models is the size of training data. Here, by (i) using training and testing examples that come from different sources and (ii) utilizing pre-trained semantic embeddings to represent sentences, we alleviated the effect of relatively small sample size to a certain extent. This approach also enabled us to explicitly test the hypothesis "use of questionnaire data in training can generate models that can be used to detect IPV on social media text with reasonable accuracy". Our results provide evidence in support of this hypothesis.

The availability of abundant data will enable training of more sophisticated machine learning models (e.g., deep learning) and likely enhance the accuracy of predictions. Such data are abundant in the social media. To this end, this study represents a significant first step towards development of such large-scale machine learning models, by providing insights into the effect of many factors, including violence types, negative examples used in training, and features that are used. Importantly, DetectIPV can also be used to generate a large corpus of text describing IPV, by iteratively applying DetectIPV to prioritize sentences and reviewing texts that are enriched in sentences scored highly by DetectIPV.

The test data that are collected from the social media may also be biased in various ways, which may influence the conclusions of this study. It is possible that people who are in abusive relationships, particularly for specific types or severity of abuse, may not post for advice publicly due to stigma and more serious red flags. This may bias the test data toward specific types of violence and/or levels of severity. In addition, red flags can be less noticeable to the victim, as physical abuse tends to be progressive over time than a sudden outburst. Thus it can be difficult to find emotional abuse in a post possibly commented in a nuanced way in the main text. IPV also has many unexplored comorbidities. In future research, our investigation of relationship dynamics that are significantly different from violent and not violent relationships can investigate the situational and context-dependent nature of IPV.

Social media data bring new contextual challenges to conducting research with high ethical standards that are different from traditional research methods [91, 92]. We followed the SoMe Report [91] to guide our ethical concerns. Specifically, the data we used are considered public data. While Facebook and closed network social media data were interpreted as private data, researchers increasingly acknowledge open discussion platforms, and open discussion platforms where people broadcast their opinions are considered as public data. However, we would like to note that the nature of our research involves sensitive data and it is important to take into consideration the sensitive nature of the data while utilizing these datasets and our models.

An important feature of DetectIPV is that it does not use data collected from individuals in training its models, i.e., the models are trained purely on questionnaires that are developed by the scientific community for research purposes. Thus, there would not be significant ethical considerations in the utilization of these models in settings where a given text is used to identify potential IPV for clinical and/or counseling purposes (i.e., when the text and the predictions are only available to the user who provides the test sentences). However, care must be taken when researchers utilize these public data to identify IPV in social media data and investigate overall trends. The social media data can also be used to train and enhance the models, which may require additional ethical considerations. To this end, consideration of privacy-preserving data mining and machine learning algorithms can be useful [93].

Conclusion

IPV can adversely effect mental, sexual, and physical health. Not being able to detect signs of red flags can danger the safety of the individuals. Early and accurate characterization of IPV can help the victims navigate their situation to relieve certain emotional, sexual, and physical consequences and guide treatment for the victims. Providing tools for exploring potential issues in their relationships is critical for improving their health and safety.

The primary outcome of this study is DetectIPV, a tool for detecting IPV from free text. DetectIPV is carefully developed and validated by taking into account multiple factors. It is available as a web service that can be used to detect IPV in query texts using pre-trained models, as well as open-source code that can be trained using other sets of training data and used as a stand-alone tool. DetectIPV can be utilized in a broad range of applications,

including flagging IPV in social media, detecting IPV in electronic health records, detecting IPV in court records, detecting IPV in transcripts of interviews, and assessing IPV in large corpuses of text for research purposes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This publication was made possible by US National Health Institutes (NIH) grant R01-LM012518 from the National Library of Medicine. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

References

1. National Center for Injury Prevention and Control, Division of Violence Prevention. (2021). Preventing intimatepartner violence fact sheet. Centers for Disease Control and Prevention, Atlanta, GA. https://www.cdc.gov/violenceprevention/pdf/ipv/IPV-factsheet_2021.pdf.
2. Breiding MJ, Basile KC, Smith SG, Black MC, Mahendra RR (2015). Intimate partner violence surveillance: uniform definitions and recommended data elements, version 2.0. National Center for Injury Prevention and Control, Centers for Disease Control and Prevention, Atlanta.
3. World Health Organization. (2013). Global and regional estimates of violence against women: Prevalence and health effects of intimate partner violence and non-partner sexual violence. World Health Organization
4. Reed LA, Tolman RM, Ward LM (2017). Gender matters: Experiences and consequences of digital dating abuse victimization in adolescent dating relationships. Journal of Adolescence, 59, 79–89 [PubMed: 28582653]
5. Velopulos CG, Carmichael H, Zakrison TL, & Crandall M. (2019). Comparison of male and female victims of intimate partner homicide and bidirectionality-an analysis of the national violent death reporting system. Journal of Trauma and Acute Care Surgery, 87(2), 331–336. [PubMed: 31348402]
6. Puzone CA, Saltzman LE, Kresnow M-J, Thompson MP, Mercy JA (2000). National trends in intimate partner homicide: United states, 1976–1995. Violence Against Women, 6(4), 409–426.
7. Afifi TO, MacMillan H, Cox BJ, Gordon J, Asmundson G, Stein MB, Sareen J. (2009). Mental health correlates of intimate partner violence in marital relationships in a nationally representative sample of males and females. Journal of Interpersonal Violence, 24(8), 1398–1417 [PubMed: 18718882]
8. Karakurt G, Patel V, Whiting K, & Koyutürk M. (2017). Mining electronic health records data: Domestic violence and adverse health effects. Journal of Family Violence, 32(1), 79–87. [PubMed: 28435184]
9. Whiting K, Liu LY, Koyutürk M, Karakurt G. (2017). Network map of adverse health effects among victims of intimate partner violence. In: Biocomputing 2017, page 324–335. WORLD SCIENTIFIC
10. Lagdon S, Armour C, & Stringer M. (2014). Adult experience of mental health outcomes as a result of intimate partner violence victimisation: A systematic review. European Journal of Psychotraumatology 5(1), 24794.
11. Hacıaliefendioğlu A, Yılmaz S, Koyutürk M, Karakurt G. (2020). Co-occurrence patterns of intimate partner violence. In BIOCOMPUTING 2021: Proceedings of the Pacific Symposium, pages 79–90. World Scientific.
12. Straus MA (1979). Measuring intrafamily conflict and violence: The conflict tactics (ct) scales. Journal of Marriage and the Family, 41(1), 75.
13. Straus MA, Hamby SL, BONEY-McCOY S, Sugarman DB (1996) The revised conflict tactics scales (cts2): Development and preliminary psychometric data. Journal of Family Issues, 17(3), 283–316.

14. Shepard M,F, Campbell J,A. (1992). The abusive behavior inventory: A measure of psychological and physical abuse. Journal of interpersonal violence, 7(3), 291–305.
15. Tolman RM (1989). The development of a measure of psychological maltreatment of women by their male partners. Violence and Victims, 4(3), 159–177. [PubMed: 2487132]
16. Smith PH, Earp JA, DeVellis R. (1995). Measuring battering: Development of the women's experience with battering (web) scale. Women's Health (Hillsdale, N.J.), 1(4), 273–288.
17. Koss M, P.,Gidycz, C. A. (1985). Sexual experiences survey: Reliability and validity. Journal of consulting and clinical psychology, 53(3), 422. [PubMed: 3874219]
18. Chu K-H, Colditz J, Malik M, Yates T, & Primack B. (2019). Identifying key target audiences for public health campaigns: Leveraging machine learning in the case of hookah tobacco smoking. Journal of Medical Internet Research, 21(7), e12443.
19. Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, Jayaraman S, Upadhaya T, & Gonzalez G. (2015). Utilizing social media data for pharmacovigilance: A review. Journal of biomedical informatics, 54, 202–212. [PubMed: 25720841]
20. Velasco E, Agheneza T, Denecke K, Kirchner G, & Eckmanns T. (2014). Social media and internet-based data in global systems for public health surveillance: A systematic review. The Milbank Quarterly, 92(1), 7–33. [PubMed: 24597553]
21. Laura LS, and Neill BB (2014). The role of facebook in crush the crave, a mobile-and social media-based smoking cessation intervention: qualitative framework analysis of posts. Journal of medical Internet research, 16(7), e3189.
22. Birnbaum M, L., Ernala S, iranmai R, Asra F, De Choudhury M, Kane JMA. (2017). Collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. Journal of Medical Internet Research, 19(8), e7956
23. Lee H-S, Lee H-R, Park J-U, & Han Y-S (2018). An abusive text detection system based on enhanced abusive and non-abusive word lists. Decision Support Systems, 113, 22–31.
24. Hammer HL (2014). Detecting threats of violence in online discussions using bigrams of important words. In 2014 IEEE Joint Intelligence and Security Informatics Conference, pages 319–319.
25. Cohen K, Johansson F, Kaati L, Mork JC (2014). Detecting linguistic markers for radical violence in social media. Terrorism and Political Violence, 26(1), 246–256
26. Warner W, Hirschberg J. (2012). Detecting hate speech on the world wide web. In Proceedings of the Second Workshop on Language in Social Media, pages 19–26, Montréal, Canada. Association for Computational Linguistics.
27. Liu S, Forss T. (2015). New classification models for detecting hate and violence web content. In 2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), volume 01, pages 487–495.
28. Garimella VRK, Alfayad A, Weber I. (2016). Social media image analysis for public health. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16, page 5543–5547. Association for Computing Machinery
29. Cer D, Yang Y, Kong S-Y, Hua N, Limtiaco N, St. John R, Constant N, Guajardo-Cespedes M, Yuan S, Tar C, Strope B, Kurzweil R. (2018). Universal sentence encoder for English. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 169–174, Brussels, Belgium, November. Association for Computational Linguistics.
30. Campbell JC (1995). Assessing dangerousness: Violence by sexual offenders, batterers, and child abusers. Sage Publications, Inc.
31. Hegarty K, Sheehan M, & Schonfeld C. (1999). A multidimensional definition of partner abuse: Development and preliminary validation of the composite abuse scale. Journal of family violence, 14(4), 399–415.
32. Hegarty K, Bush R, & Sheehan M. (2005). The composite abuse scale: further development and assessment of reliability and validity of a multidimensional partner abuse measure in clinical settings. Violence and victims, 20(5), 529–547. [PubMed: 16248489]
33. Rodenburg FA, & Fantuzzo JW (1993). The measure of wife abuse: Steps toward the development of a comprehensive assessment technique. Journal of Family Violence, 8(3), 203–228.

34. Hudson WW. (1992) Partner abuse scale: Physical (pasph). WALMYR Assessment Scales Scoring Manual. WALMYR Publishing.
35. Foshee VA, Fletcher L, G Foshee B, Karl E, Langwick SA, Arriaga XB, Heath JL, McMahon PM, & Bangdiwala S. (1996). The safe dates project: Theoretical basis, evaluation design, and selected baseline findings. American journal of preventive medicine, 12(5), 39–47.
36. Foshee VA, Bauman KE, Arriaga XB, Helms RW, Koch GG, & Linder GF (1998). An evaluation of safe dates, an adolescent dating violence prevention program. American journal of public health, 88(1), 45–50. [PubMed: 9584032]
37. Marshall LL (1992). Development of the severity of violence against women scales. Journal of family violence, 7(2), 103–121.
38. Sullivan CM., Parisian JA, Davidson WS. (1991). Index of psychological abuse: Development of a measure. In: Poster presentation at the annual conference of the American Psychological Association.
39. Sullivan CM, & Bybee DI (1999). Reducing violence using community-based advocacy for women with abusive partners. Journal of consulting and clinical psychology, 67(1), 43. [PubMed: 10028208]
40. O'Leary KD (1999). Psychological abuse: A variable deserving critical attention in domestic violence. Violence and victims, 14(1), 3–23. [PubMed: 10397623]
41. Murphy CM, Hoover SA, Taft C. (1999). The multidimensional measure of emotional abuse: Factor structure and subscale validity. In: Annual meeting of the Association for the Advancement of Behavior Therapy.
42. Sackett LA, & Saunders DG (1999). The impact of different forms of psychological abuse on battered women. Violence and victims, 14(1), 105–117. [PubMed: 10397629]
43. Tolman RM (1999). The validation of the psychological maltreatment of women inventory. Violence and victims, 14(1), 25–37. [PubMed: 10397624]
44. Smith PH, Smith JB, & Earp JAL (1999). Beyond the measurement trap: A reconstructed conceptualization and measurement of woman battering. Psychology of Women Quarterly, 23(1), 177–193.
45. Smith PH, Thornton GE, DeVellis R, Earp J, & Coker AL (2002). A population-based study of the prevalence and distinctiveness of battering, physical assault, and sexual assault in intimate relationships. Violence against women, 8(10), 1208–1232.
46. Kilpatrick D, Edmunds C, & Seymour AK (1992). The national women's study. National Victim Center.
47. Resnick HS, Kilpatrick DG, Dansky BS, Saunders BE, & Best CL (1993). Prevalence of civilian trauma and posttraumatic stress disorder in a representative national sample of women. Journal of consulting and clinical psychology, 61(6), 984. [PubMed: 8113499]
48. Tjaden P. & Thoennes N. (2000). Full report of the prevalence, incidence, and consequences of violence against women: Findings from the national violence against women survey. Annotation.
49. Koss MP, & Oros CJ (1982). Sexual experiences survey: A research instrument investigating sexual aggression and victimization. Journal of Consulting Psychology, 50(3), 455–457.
50. Koss MP, Gidycz CA, & Wisniewski N. (1987). The scope of rape: Incidence and prevalence of sexual aggression and victimization in a national sample of higher education students. Journal of consulting and clinical psychology, 55(2), 162. [PubMed: 3494755]
51. Belknap J, Fisher BS, & Cullen FT (1999). The development of a comprehensive measure of the sexual victimization of college women. Violence Against Women, 5(2), 185–214.
52. Fisher BS, Cullen FT, Turner MG (2000). The sexual victimization of college women. Research Report.
53. Busby DM, Christensen C, Crane DR, & Larson JH. (1995). A revision of the dyadic adjustment scale for use with distressed and nondistressed couples: Construct hierarchy and multidimensional scales. Journal of Marital and family Therapy, 21(3), 289–308.
54. Endler NS, & Parker J. (1990). Coping inventory for stressful situations. Multi-Health systems Incorporated.

55. Brennan KA, Clark CL, & Shaver PR (1998). Self-report measurement of adult attachment: Anintegrative overview. In Simpson JA & Rholes WS (Eds.), Attachment theory and close relationships (pp. 46–76). The Guilford Press.
56. Hamby SL (1996). The dominance scale: Preliminary psychometric properties. Violence and Victims, 11(3), 199–212. [PubMed: 9125789]
57. Ware JE Jr., Sherbourne CD (1992) The MOS 36-item short-form health survey (SF-36). I. Conceptual framework anditem selection. Medical Care, 30(6), 473–483. [PubMed: 1593914]
58. Bodenmann G. (2008). Dyadic coping and the significance of this concept for prevention and therapy. Zeitschrift für Gesundheitspsychologie, 16(3), 108–111.
59. Erickson RJ (1993). Reconceptualizing family work: The effect of emotion work on perceptions of marital quality. Journal of Marriage and the Family, 55(4), 888–900.
60. Elliott DM, & Briere J. (1992). Sexual abuse trauma among professional women: Validating the trauma symptom checklist-40 (tsc-40). Child abuse & Neglect, 16(3), 391–398. [PubMed: 1617473]
61. Watson D, Clark LA, & Tellegen A. (1988). Development and validation of brief measures of positive and negative affect: The panas scales. Journal of personality and social psychology, 54(6), 1063. [PubMed: 3397865]
62. Davis MH (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. Journal of personality and social psychology, 44(1), 113.
63. Lambert MJ, Burlingame GM, Umphress V, Hansen NB, Vermeersch DA, Clouse GC, & Yanchar SC (1996). The reliability and validity of the outcome questionnaire. Clinical Psychology & Psychotherapy: An International Journal of Theory and Practice, 3(4), 249–258.
64. Andrews P, & Meyer RG (2003). Marlowe–crowne social desirability scale and short form C: Forensic norms. Journal of clinical psychology, 59(4), 483–492. [PubMed: 12652639]
65. John OP, Srivastava S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives, vol 2. University of California Berkeley.
66. Marshall GN, & Hays RD (1994). The patient satisfaction questionnaire short-form (PSQ-18), vol 7865. Rand Santa Monica.
67. Avolio BJ, Bass BM, & Jung DI (1999). Re-examining the components of transformational and transactional leadership using the multifactor leadership. Journal of occupational and organizational psychology, 72(4), 441–462.
68. Howlett A. (2015). Predictors of academic achievement, motivation and student disengagement in university students. PhD thesis, University of Tasmania.
69. Lam S-F., Jimerson S, Wong BPH, Kikas E, Shin H, Veiga FH, Hatzichristou C, Polychroni F, Cefai C, & Negovan. (2014). Understanding and measuring student engagement in school: The results of an international study from 12 countries. School Psychology Quarterly, 29(2), 213. [PubMed: 24933218]
70. Jorgensen BL(2007) Financial literacy of college students: Parental and peer influences. PhD thesis, Virginia Tech.
71. Yiyun P, & Linda NB (2015). Driver's adaptive glance behavior to in-vehicle information systems. Accident Analysis & Prevention, 85, 93–101. [PubMed: 26406538]
72. Ersche KD, Lim T-V, Ward LHE, Robbins TW, Jan, & S. (2017). Creature of habit: A self-report measure of habitual routines and automatic tendencies in everyday life. Personality and Individual Differences, 116 73–85. [PubMed: 28974825]
73. British Health Foundation. Lifestyle Questionaire.
74. Taylor HL, Jacobs DR Jr, Schucker B, Knudsen J, Leon AS, & Debacker G. (1978). A questionnaire for the assessment of leisure time physical activities. Journal of chronic diseases, 31(12), 741–755. [PubMed: 748370]
75. Pianta RC, & Steinberg M. (1992). Teacher–child relationships and theprocess of adjusting to school. In Pianta RC (Ed.), Beyond the parent: The role of other adults in children's lives (pp. 61–80). Jossey-Bass.
76. Troyer AK, & Rich JB (2002). Psychometric properties of a new metamemory questionnaire for older adults. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 57(1), P19–P27. [PubMed: 11773220]

77. Northouse PG (2014). Leadership: Theory and practice. Sage publications.
78. Thompson MP, Basile KC, Hertz MF, Sitterle D. (2006). Measuringintimate partner violence and victimization and perpetration: A compendium of assessment tools. Centers for Disease Control and Prevention, Atlanta, GA. http://www.cdc.gov/ncipc/pub-res/IPV_Compendium.pdf.
79. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems, vol 26. Curran Associates, Inc.
80. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM (2019). ChestX-ray: Hospital-Scale Chest X-ray Database and Benchmarks on Weakly Supervised Classification and Localization of Common Thorax Diseases, p 369–392. Advances in Computer Vision and Pattern Recognition. Springer International Publishing.
81. Chen Q, Yifan P, Zhiyong L. (2020). Biosentvec: Creating sentence embeddings for biomedical texts. arXiv.
82. Majumder SB, & Das D. (2020). Detecting fake news spreaders on twitter using universal sentence encoder. In: CLEF (Working Notes).
83. Asgari-Chenaghlu M, Nikzad-Khasmakhi N, Minaee S. (2020). Covid-transformer: Detecting trending topics on twitter using universal sentence encoder. arXiv e-prints, pages arXiv–2009.
84. Iyyer M, Manjunatha V, Boyd-Graber J, Hal D III. (2015). Deep unordered composition rivals syntactic methods for text classification. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), p 1681–1691, Beijing, China. Association for Computational Linguistics.
85. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. (2017). Attention is all you need. In Advances in neural information processing systems, p 5998–6008.
86. Ríssola EA, Losada DE Crestani, F. (2021). A survey of computational methods for online mental state assessment on social media. ACM Transactions on Computing for Healthcare, 2, 1–31. 10.1145/3437259.
87. Davis A. (2014). Violence-related mild traumatic brain injury in women: Identifying a triad of postinjury disorders. Journal of Trauma Nursing|, 21(6), 300–308. [PubMed: 25397339]
88. Christ C, De Waal MM, Dekker JJM, van Kuijk I, Van Schaik DJF, Kikkert MJ, & Messman-Moore TL. (2019). Linking childhood emotional abuse and depressive symptoms: The role of emotion dysregulation and interpersonal problems. PLoS One, 14(2). e0211882.
89. Follingstad DR (2009). The impact of psychological aggression on women's mental health and behavior: The status of the field. Trauma, Violence, & Abuse, 10(3), 271–289.
90. Engel B. (2002). The emotionally abusive relationship: How to stop being abused and how to stop abusing. John Wiley & Sons.
91. Hunter RF, Gough A, O'Kane N, McKeown G, Fitzpatrick A, Walker T, McKinley M, Lee M, & Kee F. (2018). Ethical issues in social media research for public health. American Journal of Public Health, 108(3):343–348. [PubMed: 29346005]
92. Townsend L, & Wallace C. (2016). Social media research: A guide to ethics. University of Aberdeen, 1, 16.
93. Al-Rubaie M, & Chang J M. (2019). Privacy-preserving machine learning: Threats and solutions. IEEE Security & Privacy, 17(2), 49–58.

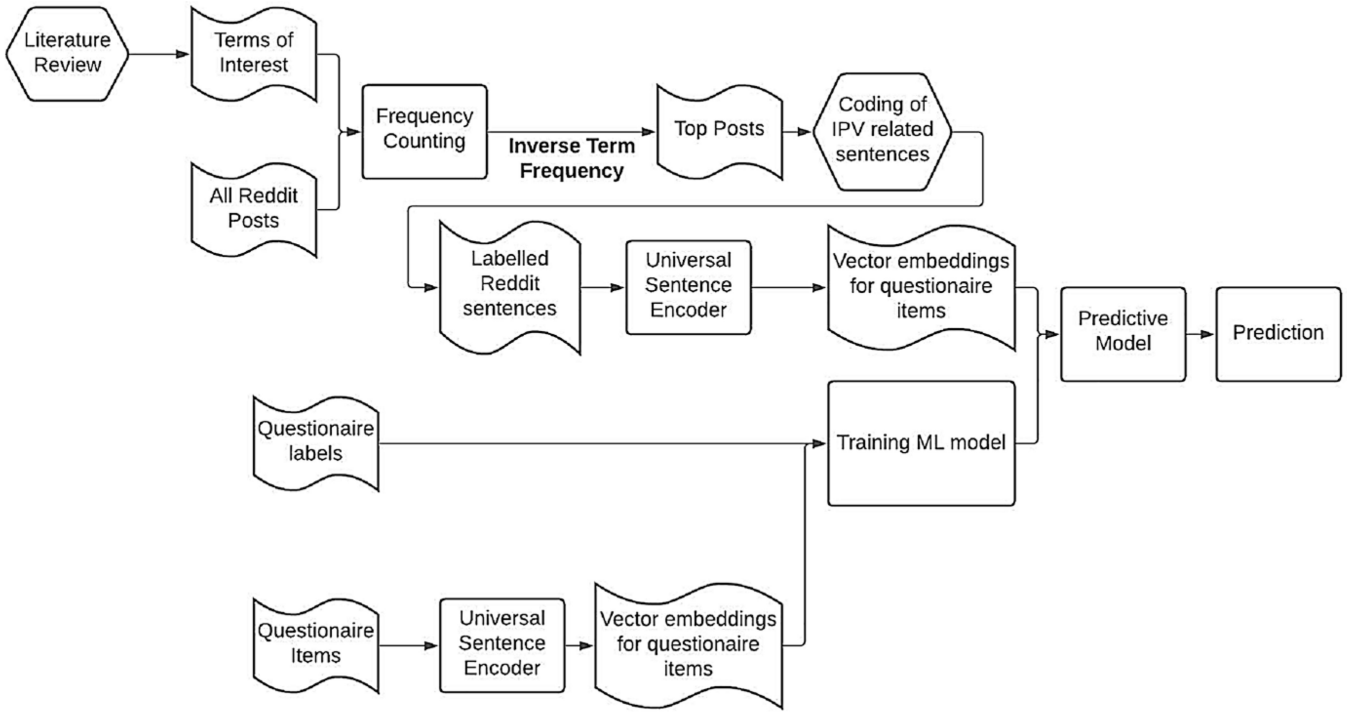

Fig. 1.

The workflow of the machine learning framework for predicting IPV from social media posts. Boxes with curved sides show data (input, intermediate, or output), honeycombs show human input (expert review, literature search coding), and rectangles show algorithms/tools/processing

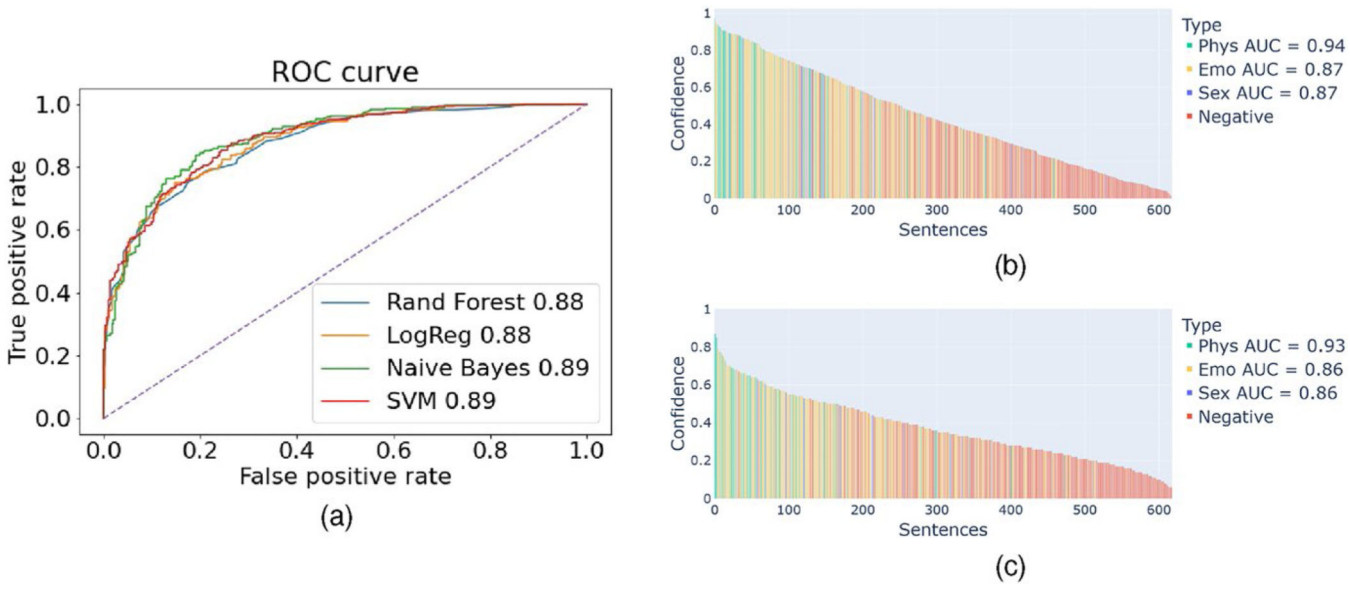

Fig. 2.

Predictive accuracy of the “General Violence Model” using different machine learning algorithms. **a** The ROC curve for the predictions provided by four different machine learning algorithms on the testing (social media) data, where sentences are labeled “positive” or “negative” depending on the presence of violence in the sentence. **b** The predictions of the Logistic Regression model on the testing (social media) data, where sentences are ordered according to confidence of predictions and colors show sentence label such that positive samples are annotated with type of violence. **c** The predictions of the Random Forest model on the testing (social media) data. *Phys* physical abuse, *Emo* emotional abuse, *Sex* sexual abuse, *Negative* no violence. Each violence type is indicated with Area Under ROC, or Area Under Curve (AUC)

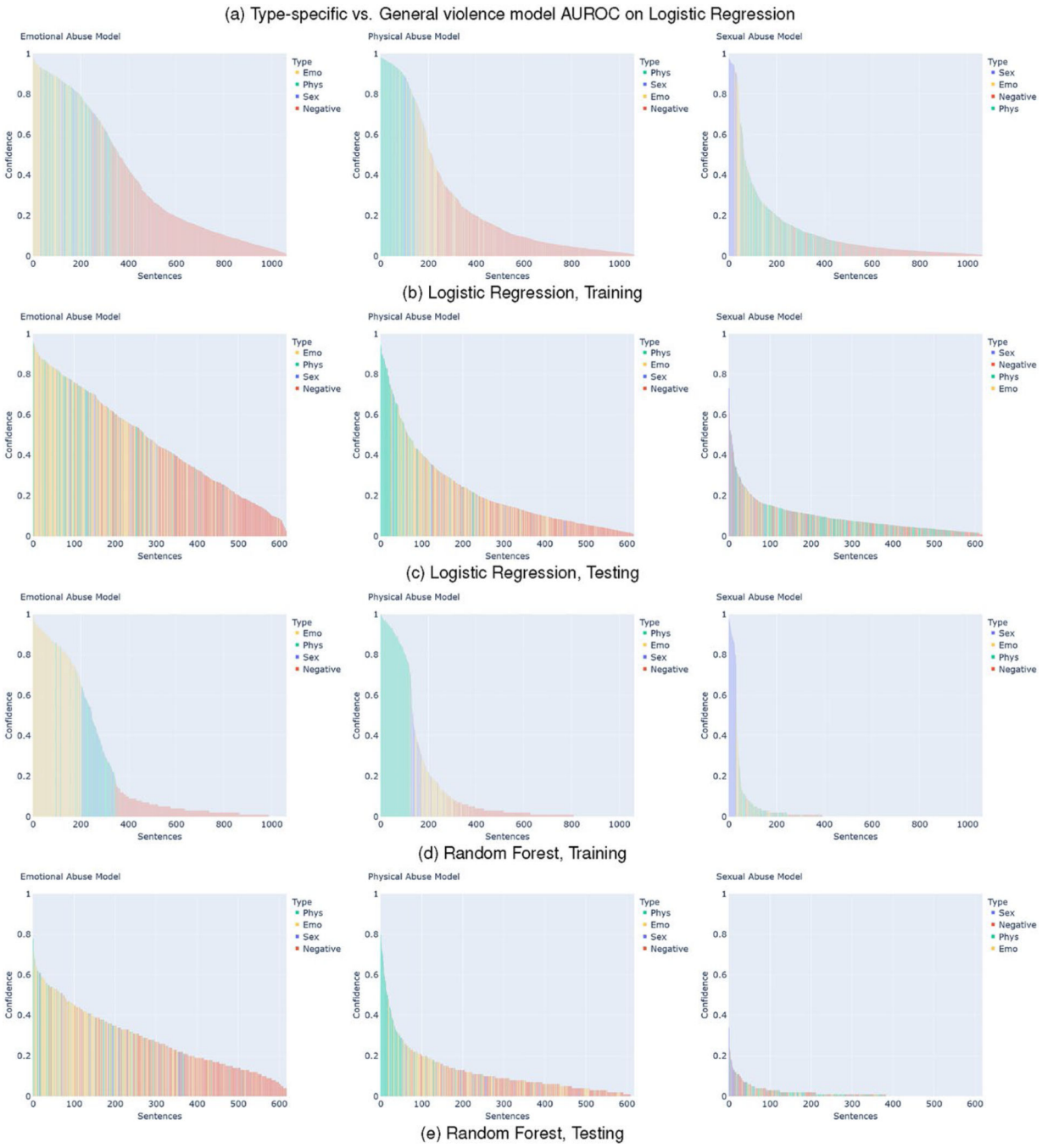

Fig. 3.

Predictive performance of type-specific violence models. The type-specific models are trained by considering the specified IPV types as positive samples and sentences with no violence as negative samples. For a type-specific model, positive samples representing the other two types are not included in the training data. **a** The comparison of the performance of the type-specific violence model in predicting violence against the general violence model based on all type-specific sentences and negatives in the test (social media) data (type-specific (left) vs general (right)). Positive test samples representing the other two types

are not considered in computing the ROC. **b**–**e** Each row shows the waterfall plot of the confidence scores assigned by the type-specific model to all sentences in the (**b**, **d**) training (questionnaire) and (**c**, **e**) test (social media) data for emotional, physical, and sexual abuse models from left to right. The bars are colored according to violence type (or no violence/ negative). **b**, **c** Logistic Regression, (**d**, **e**) Random Forest

Fig. 4.

Radial visualization of the predictions of type-specific models for training and testing sentences. Upper panel: Random Forest model, Lower panel: Logistic Regression model; Left: Training (Questionnaire) Sentences, Right: Testing (Social Media) Sentences. Each axis represents an IPV type (0°: emotional abuse, 120°: sexual abuse, 240°: physical abuse). The distance of a point from the center represents the magnitude of the degree of confidence assigned by the three type-specific models to the corresponding sentence, while its angle/direction represents the dominant violence type in terms of the confidence assigned by the models. The color of the point represents the true label assigned to each sentence by expert reviewers, which can include of multiple types of IPV, as shown by the color bar.

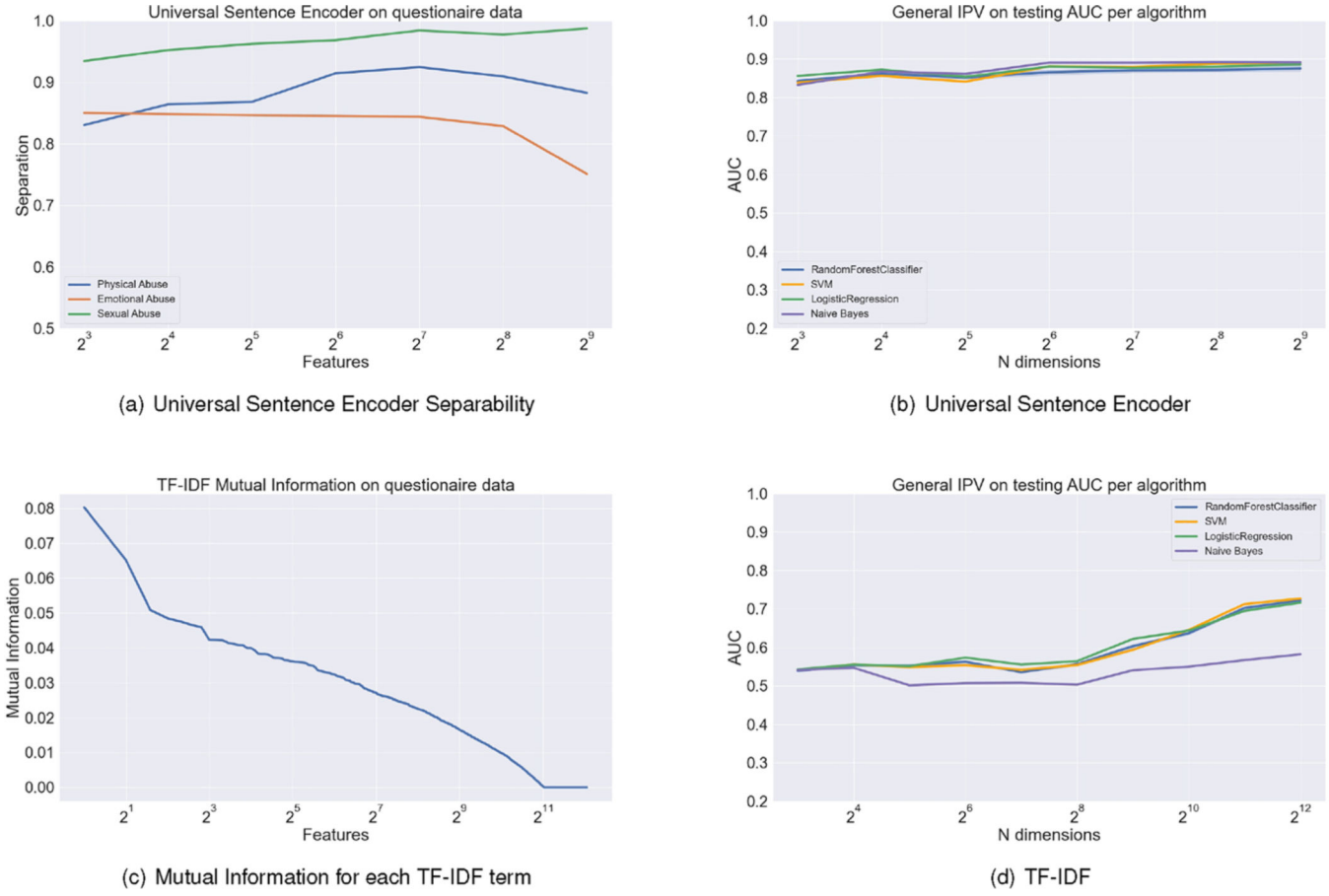

Fig. 5.

Comparison between Universal Sentence Encoder (USE) vs. term-frequency (TF-IDF) based prediction of IPV. Left: Feature prioritization and selection. For USE, dimensions are ordered according to separability of general violence, cumulative separability of each violence type is shown in the plots (**a**). For term-frequency, the terms are ordered according to mutual information with sentence labels and rank vs. mutual information is plotted in (**c**). Right: The predictive performance (AUROC) of the general violence models on the test (social media) sentences as a function of number of features using USE (b) and TF-IDF (**d**). The features are selected based on the ordering on the left (computed on training sentences) and four different models are trained for each number of features using four different classification algorithms

Table 1

Sample sentences taken from social media (Reddit)

Physical abuse	• Screaming, throwing things, punching holes in doors and walls, smashing my phone with a baseball bat • He has shoved me on numerous occasions and squeezed my fingers, arm, legs to the point of pain and pinned me physically or blocked me from leaving our home. • He has choked me, kicked me, pulled my hair a couple times. • He grabbed my face very tight and shook it hard, I got scared and shook him off me and locked myself in the bedroom. • I was crying on the bed, fingers in my ears humming to stop from hearing him berate me - because he wasn't letting me leave, he'd push me back onto the bed if I tried to get up, take my phone away from me, (I get scared when a 117kg, 6"3 man whos explosive is yelling and punching sh*t, I have PTSD from when he used to do it a lot more) he'd be pacing the room saying horrible sh*t to me. • The moment that really scared me was when he grabbed me with both hands around my neck.
Emotional abuse	• I had convinced myself that I was disgusting and undeserving of love and she was just trying to help, but I now realise she probably wasn't even trying to help me at all. • I realised there were a lot of things in our relationship that I was uncomfortable with that I felt like I could never bring up because she would get upset and start crying so I would end up comforting her, or she would just turn it on me and get angry and blame me. • She locked me out while I was just in my socks with my wallet still in the house. • He says that his ignoring me is my fault because of the way I'm acting. • Denying everything, telling me I'm stup*d, and one time calling me a b*tch.
Sexual abuse	• He spends a lot of time obsessively badgering me and asking for intimate details of my past relationship, which he then uses against me and he constantly treats me with suspicion and accuses me of lying about my sexual history with my ex (which was brief and very limited). • He wants me to dress this way... sit that way... blow him this way... s*ck him that way... Hair up... High pony-tail (if you want me to be sexually attracted to you, and you want me to be interested in sex -I'll tell you what turns me on... he says).... I asks for small things..... small..... he goes limp... loses interest completely. • How am I a narcissist I'm just saying don't f*ck me bc it hurts.

Left label, *Right* sample sentences

Table 2

Questionnaires and scales used as training examples

Physical violence	Emotional violence	Sexual violence
Abusive Behavior Inventory [30] Composite Abuse Scale [31, 32] Measure of Wife Abuse [33] Partner Abuse Scale - Physical [34] Revised Conflict Tactic Scale [13] Safe Dates - Physical Violence Victimization [35, 36] Severity of Violence Against Women Scale (SVAWS) [37]	Abusive Behavior Inventory [14] Composite Abuse Scale[31, 32] Index of Psychological Abuse [38–40] Measure of Wife Abuse [33] Multidimensional Measure of Emotional Abuse [41] Partner Abuse Scale - Non-physical [42] Psychological Maltreatment of Women Inventory (PMWI) [15, 43] Psychological Maltreatment of Women Inventory (PMWI)-Short Form [43] Revised Conflict Tactics Scales (CTS-2) [13] Safe Dates - Psychological Abuse Victimization [35, 36] Women's Experiences with Battering (WEB) [16, 44] [45]	Measure of Wife Abuse [33] National Women's Study (NWS) and National Violence Against Women Survey (NVAWS) [46–48] Revised Conflict Tactic Scale [13] Severity of Violence Against Women Scale (SVAWS) [37] Sexual Experiences Survey (SES) - Victimization Version [49, 17, 50] Sexual Victimization of College Women [51,52]
Related negatives	Unrelated negatives	
Revised Dyadic Adjustment Scale [53] Coping Inventory for Stressful Situations [54] Experiences in Close Relationships Scale [55] Dominance Scale [56] Short Form Health Survey (SF-36) [57] Dyadic Coping Inventory [58] Marital Burnout [59] Trauma Symptom Checklist[60] The Positive and Negative Affect Schedule [61] Interpersonal Reactivity Index [62] Brief Symptom Inventory [63]	Marlowe-Crowne Social Desirability Scale 13-Item Short Form [64] The Big Five Inventory Personality Test [65] The Patient Satisfaction Questionnaire Short Form [66] Multifactor Leadership Questionnaire [67] Academic Progress Questionnaire [68] Student Engagement in Schools Questionnaire [69] College Students Financial Literacy Survey (CSFLS) [70] Driver's adaptive glance behavior [71] Creature of Habit Scale [72] Lifestyle Questionnaire (British Heart Foundation)[73] Assessment of Leisure Time Physical Activities [74] Student Teacher Relationship Survey [75] Multifactorial Memory Questionnaire [76] Conceptualizing Leadership Questionnaire [77]	

Top positive examples, *Bottom* negative examples

Table 3

Training and testing data used in our computational experiments

	Positive Examples				Negative Examples		
Dataset	**Physical**	**Emotional**	**Sexual**	**Total**	**Related**	**Unrelated**	**Total**
Training: questionnaire Items	126	198	31	355	297	410	707
Testing: sentences from Reddit Posts	74	257	25	326	96	196	292

Each entry shows the number of positive and negative examples of the respective type in the training and testing data. The total number of positive examples is less than the sum of the number of examples in each violence type, since a single sentence can be annotated with multiple violence types

Table 4

The effect of using relationship-related or unrelated negative sentences in training and testing on predictive performance

(a) AUROC on questionnaire data for 5 runs of 5-fold cross validation						
Negative Type	**Validation Unrelated**		**Validation Related**		**Validation Both**	
Training Unrelated	0.97 ±0.03	0.97 ±0.04	0.85 ±0.10	0.91 ± 0.04	0.92 ± 0.04	0.95 ± 0.04
	0.97 ±0.03	0.97 ±0.03	0.83 ±0.11	0.90 ± 0.05	0.91 ± 0.04	0.94 ± 0.03
Training Related	0.94 ± 0.05	0.96 ± 0.04	0.95 ± 0.02	0.97 ± 0.03	0.95 ± 0.03	0.96 ± 0.03
	0.96 ±0.03	0.95 ± 0.04	0.92 ± 0.04	0.95 ± 0.03	0.94 ± 0.02	0.95 ± 0.03
Training Both	0.97 ±0.03	0.97 ±0.04	0.93 ± 0.04	0.96 ± 0.03	0.95 ± 0.02	0.97 ±0.03
	0.97 ±0.02	0.97 ±0.03	0.88 ± 0.08	0.94 ± 0.03	0.93 ± 0.03	0.96 ± 0.02

(b) AUROC on social media data for the model trained using all questionnaire data						
Negative Type	**Testing Unrelated**		**Testing Related**		**Testing Both**	
Training Unrelated	0.91	0.93	0.77	0.81	0.86	0.89
	0.93	0.92	0.78	0.81	0.88	0.88
Training Related	0.83	0.89	0.78	0.82	0.82	0.86
	0.90	0.87	0.79	0.80	0.86	0.85
Training Both	0.91	0.91	0.83	0.84	0.88	0.89
	0.94	0.91	0.80	0.82	0.89	0.88

Each entry shows the area under ROC curve for four machine learning algorithms on a task obtained using relationship-related negatives, unrelated negatives, and both types of negatives in training vs. testing. The numbers for each machine learning algorithm is shown in a different color: Random Forest (cyan), *SVM* (orange), *Naive Bayes* (purple), **`Logistic Regression`** (olive)