Using machine learning algorithms (supervised) to generate automatically labeled dataset for detecting digital dating abuse from text messages

Tania Roy,[1] Thomas Maranzatto,[2] Zachary Loomas[1]
[1]New College of Florida
[2]University of Illinois at Chicago
troy@ncf.edu, tmaran2@uic.edu, zachary.loomas12@ncf.edu

Abstract

Digital dating abuse is a form of intimate partner violence that uses technology as a medium to propagate fear and cause harm for dating partners. Over several years digital dating abuse has been on the rise, and particularly during COVID-19, the issue has risen exponentially. This project aims to create a tool that raises awareness and detects digital dating from text messages. Previously, we generated a dataset with expert labelers to use supervised machine learning algorithms for abuse detection. However, the cost and time associated with generating human-annotated datasets limit the size of these verified datasets. This poster explores using machine learning algorithms trained on human-annotated datasets to label more extensive crowd-sourced datasets and generate a larger training dataset for abuse detection algorithms. We used Naive Bayes, Decision Tree, LSVM, and LSTM to test for accuracy and speed of labeling this more extensive dataset.

Introduction

Communication amongst young and older adults, particularly during the pandemic, has shifted to an online medium. The use of email, text, and instant messaging services is prevalent. Unfortunately, these internet services can also harass people indiscriminately (Stonard et al. 2014). One such form of harassment is digital dating abuse, where a dating partner leaves repeated threatening messages over the phone or on social media. Studies have shown that undergraduate college students are prone to this aggression. One-third (1/3) of college students reported having faced cyber harassment (Spitzberg and Cupach 2014). The focus group studies by Melander (Melander 2010) identified themes relating to using technology to abuse someone. Participants also said in-person, aggressive, and private arguments could quickly become aggressive public-domain conversations (Carlson 2003; Roy et al. 2016). Digital dating abuse has become a significant mental health crisis for young adults.

Related Work

De Chowdhury et al. administered standardized tests to determine if users were depressed. They used 476 ”good” participants and found that 171 had depression scores above the threshold. SVM and RBF were used to classify the feature space (De Choudhury et al. 2013). In his work on analyzing domestic abuse on social media data, Schrading used Reddit and Twitter data to show that a classifier can detect a reason for leaving or staying in an abusive relationship. The researchers used the perceptron algorithm, SVM, and neural network classifiers (NN). Linear Support Vector Machine achieved a 90% accuracy rate.(Schrading et al. 2015).

Our research on digital dating abuse is closely related to the work mentioned above (Roy et al. 2016; Roy, Young, and Hodges 2020). As digital dating abuse falls under interpersonal violence, we predicted that we could use similar machine-learning methods to conduct our studies. In prior work, we (Roy, McClendon, and Hodges 2018) created a detection application that combines an LSVM, tf-idf feature extractor, and unigram input to label text messages as abusive vs. non-abusive in the context of digital dating abuse with an accuracy of 91.4%. For the training and testing dataset, 161 abusive text messages are part of this dataset. We also created three additional balanced datasets with these abusive and non-abusive text messages from two publicly available datasets (SPAM, Mobile Forensics Text Message Corpus, and a combination of the above two) (Cormack, Hidalgo, and Sánz 2007; O'Day and Calix 2013).

Goals

We had two primary goals with this poster: 1) Provide a new dating abuse dataset based on real-world experiences of dating abuse survivors. 2) Provide preliminary results and evaluate robust machine learning-based techniques to classify this new dataset given limited training and testing data.

Methodology

Dataset Collection

For this study, we recruited 135 participants through Mechanical Turk who responded to 10 prompts primed to invoke negative sentiment. The prompts were structured so participants would answer in a text message style and be all related to dating abuse scenarios. The goal of the current dataset was to overcome some of the limitations of the older dataset created by us in 2018 (Roy, McClendon, and Hodges 2018).

Dataset Creation

They were presented with an abusive scenario, such as

> *Please read the following scenario carefully:* I really like this guy, but we are both 15! He is taking our relationship too far. He keeps messaging me with embarrassing details about him and pushes to have sex. I keep saying no. *Please provide three text messages that her boyfriend might have sent her that would be over the line.*

After removing blank responses, the researchers had a raw dataset containing 1350 messages which yielded 1200 final text messages. Some message entities contain multiple lines of text, and some are single lines resulting in several thousand responses. Due to the volume of the text messages, we found using human annotators who are domain experts time intensive and expensive. Another consideration would be the mental health and emotional impact of these text messages in large volumes on the annotators. Thus we turn to supervised machine learning techniques to label messages automatically.

Our goal with this dataset is to be able to classify each response as abusive or non-abusive. Because of the complexity of the data, before classification, each message in our dataset had to be cleaned. We began this process by making all messages lowercase. We defined a dictionary of 116 common contractions and associated an expansion for each. This expansion does not map directly to the meaning in a given sentence. For example, we define the map he'd $\rightarrow$ he would, but another valid expansion would be he'd $\rightarrow$ he had. Messages were then tokenized (splitting a sentence into words) and stemmed (prefixes and suffixes were removed) each message. Finally, we extracted the features of each message using unigram tf-idf vectorization. We call this final cleaned dataset **d2019**. We follow this same procedure on the dataset from Roy et al. (2018) (Roy, McClendon, and Hodges 2018), and call this dataset **d2018**. Finally, a third dataset, UCI's Ham/ Spam corpus (Cormack, Hidalgo, and Sánz 2007), is cleaned analogously, and we call this dataset **dHam**.

Training and Testing Models

We used d2018 as a training/testing sample for our dataset. The training set had 113 abusive and 100 non-abusive messages, while the testing set had 48 abusive and 40 non-abusive messages. We trained three supervised models on this data: We chose these models as they gave favorable results for our previous work. (Roy, McClendon, and Hodges 2018). These three models were then used to classify responses obtained from the survey and the "ham" messages (dHam). These three classifiers label each message based on the majority vote. We used this ensemble methodology to reduce bias any classifier might project onto the data as our training set was relatively small.

We then trained and evaluated nine classifiers with these newly labeled messages. In all cases, a Linear SVM with regularization = 0.1, a decision tree, and a multinomial naive Bayes were trained from the new data (Pedregosa et al. 2011). These nine classifiers were evaluated on Roy et.al's 2018 test dataset(Roy, McClendon, and Hodges 2018).

Evaluation

To evaluate the performance of our models, we used four performance metrics: accuracy, confusion matrices, F1-Score, and ROC AUC score(Kohavi 1995),(Hanley and McNeil 1982 04),(Ramos 2003). The Naive Bayes, Decision Tree, SVC, and feedforward neural network were tested using unigram, bigram, and trigram inputs. The linear support vector machine consistently performed better than other classifiers across all three datasets. The best mean accuracy was obtained with a linear SVC on the combined dataset with a value of 0.964. The values for the d2019 and HAM datasets are 0.902 and 0.878, respectively.

The best accuracy for the unseen test set (obtained from d2018) was consistently the linear SVC. Again the combined dataset provided the best results with a value of 0.931. We next wanted to evaluate the misclassifications of each model on the d2018 test set using confusion matrices. We are interested in confusion matrices with balanced false positives and false negatives classes. The best possible confusion matrix for the unseen dataset is [48,0][0,40]. From the raw accuracy results, we expect the SVC classifier on the combined training dataset to perform the best. Only six messages were misclassified, as observed in the confusion matrix results. The SVC on the HAM dataset performed worse but had a better spread of false positives and false negatives.

The LSTM model performed poorly on the HAM dataset, with an accuracy worse than any unigram classifier. Trained on the 2019 and Combined datasets, the LSTM performed better with 87.5% and 89.8% accuracy, respectively. However, given the relative complexity of the LSTM classifier compared to the SVC or Naive Bayes classifiers, this performance could be better. We used the F1 metric and ROC AUC scores to prove that the combined dataset and the SVC provide the most predicting power for unseen messages (0.934). Again, the SVC with the combined dataset performs the best for the ROC AUC scores, with a score of 0.912. Across all classifiers, the combined dataset gave the best results.

Limitation

Limitations of this methodology include the following: the performance of the models depends directly on the size and accuracy of professionally annotated dating abuse messages. A larger corpus of annotated messages means machine models can learn complex edge case messages more reliably; our initial d2018 dataset was small. Binary classification needs to address the ambiguities of the grey area, which is crucial for this topic. This work aims to power a mobile phone app's detection tool to help users detect and get resources for help. This work is a step forward in creating a more diverse and inclusive dataset for the app, reducing the chances of incorrect and biased detection.

References

Carlson, C. N. 2003. Invisible victims: Holding the educational system liable for teen dating violence at school. 26:351.

Cormack, G. V.; Hidalgo, J. M. G.; and Sánz, E. P. 2007.

Feature engineering for mobile (SMS) spam filtering. 871–872. ACM.

De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting depression via social media. 2.

Hanley, J. A., and McNeil, B. J. 1982-04. The meaning and use of the area under a receiver operating characteristic (ROC) curve. 143(1):29–36.

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. volume 14, 1137–1145. Stanford, CA.

Melander, L. A. 2010. College students' perceptions of intimate partner cyber harassment. 13(3):263–268.

O'Day, D. R., and Calix, R. A. 2013. Text message corpus: applying natural language processing to mobile device forensics. 1–6. IEEE.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; and Dubourg, V. 2011. Scikit-learn: Machine learning in python. 12:2825–2830.

Ramos, J. 2003. Using tf-idf to determine word relevance in document queries.

Roy, T.; Hodges, L. F.; Daily, S. B.; and McClendon, J. 2016. Secondlook: Participatory design process to create a phone app that detects digital dating abuse. In *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, 320–327.

Roy, T.; McClendon, J.; and Hodges, L. 2018. Analyzing abusive text messages to detect digital dating abuse.

Roy, T.; Young, E.; and Hodges, L. F. 2020. A second look at secondlook: Design iterations and usability of digital dating abuse detection and awareness app. In *2020 IEEE International Conference on Healthcare Informatics (ICHI)*, 1–11.

Schrading, N.; Alm, C. O.; Ptucha, R. W.; and Homan, C. 2015. # WhyIStayed,# WhyILeft: Microblogging to make sense of domestic abuse. 1281–1286.

Spitzberg, B. H., and Cupach, W. R. 2014. *The dark side of relationship pursuit: From attraction to obsession and stalking*. Routledge.

Stonard, K. E.; Bowen, E.; Lawrence, T. R.; and Price, S. A. 2014. The relevance of technology to the nature, prevalence and impact of adolescent dating violence and abuse: A research synthesis. 19(4):390–417.