Silent Signals: Identifying Abuse Through Social Media Content

Yonathan Afek, Gefen Pustelnic, Itamar Shashar

TECHNION Israel Institute of Technology

פורום מיכל סלה Michal Sela Forum

Introduction

- **Abusive relationships** are a critical issue that affect people worldwide.
- With the rise of **social media**, more people are sharing personal stories.
- Using **NLP**, we can **detect signs of abuse** and **identify key risk factors.**
- Allows for efficient identification of **abusive patterns** and better **support for victims**.

Data

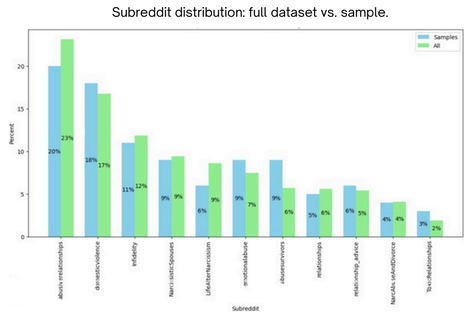
Dataset of 10k Reddit posts

- Post Metadata
- Relationship and Demographic data
- Contexual Risk Factors

u/NoChoice682 · 23 hr ago ·
What stopped you from leaving your abusive partner?
I had no support system

Data Validation

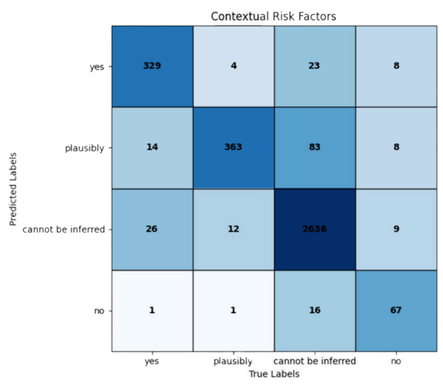
- **Sampling** 100 examples.
- **Verifying** that the data is representative.
- **Manually classifying** based on the stories.



Data Exploration

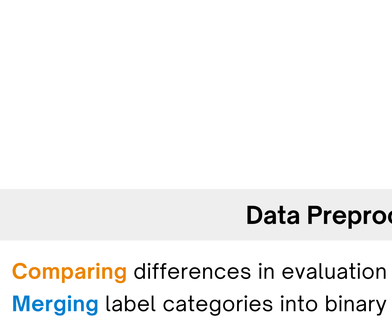
Insights:

- **Biased** towards abuse classification.
- **Relied** on author's descriptions.
- **Misinterpreted** victim traits as offender.
- **Failed** to infer social norms.
- **Assumed** the victim is female.



Data Preprocessing for models' prediction

- Comparing differences in evaluation metrics.
- Merging label categories into binary classifications.
- Selecting the most reliable risk factors for the following parts.



Risk Factor	F1 - new	F1 - old	Difference
ptsd	1.0	0.82	0.18
substance_use	1.0	1.0	0.0
access_to_weapons	1.0	1.0	0.0
economic_violence	1.0	0.83	0.17
signs_of_injury	0.97	0.94	0.03
emotional_dependency	0.97	0.65	0.32
physical_violence	0.96	0.89	0.07
property_damage	0.96	0.91	0.05
social_isolation	0.96	0.96	0.0
aggressive_behavior	0.96	0.8	0.16
attempts_to_end_relationship	0.95	0.86	0.09
gaslighting	0.95	0.93	0.02
sexual_violence	0.94	0.93	0.01
mental_condition	0.94	0.91	0.03
daily_activity_control	0.94	0.88	0.06
refusal_to_end_relationship	0.93	0.88	0.05
public_private_discrepancy	0.93	0.91	0.02
presence_of_others_in_assault	0.93	0.88	0.05
outbursts	0.93	0.91	0.02
narcissistic_traits	0.92	0.88	0.04
humiliation	0.92	0.82	0.1
living_with	0.91	0.91	0.0

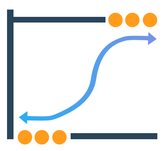
Baseline Models

Goals:

1. Establish a baseline for score.
2. Data Insights: Identify key features and important words.

We predicted the risk factor: **physical violence.**

Logistic Regression

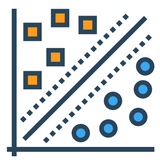


First try: ***F1 = 0.86, accuracy = 0.88***

- Removing risk factors with coefficients having p-values > 0.05

Final result: ***F1 = 0.86, accuracy = 0.89, Pseudo R^2 = 0.572***

tf-idf with SVM classifier



- **Embedding**: Vector of the 1000 most important words using TF-IDF.
- **Comparison**: Applied an SVM classifier with different kernels.
- **Result**: The RBF kernel achieved **F1 score** of **0.79**.

Keywords detection for each risk factor:

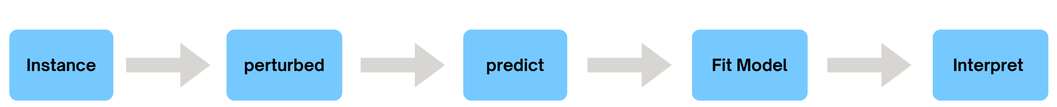
Top 10 terms for physical violence:		
	Term	TF-IDF Ratio
380	mum	71.660
253	cat	57.409
344	proof	49.271
641	judge	48.748
759	spouse	48.048
742	power	47.788
124	black	46.725
590	neck	46.502
753	death	46.167
278	legal	45.004

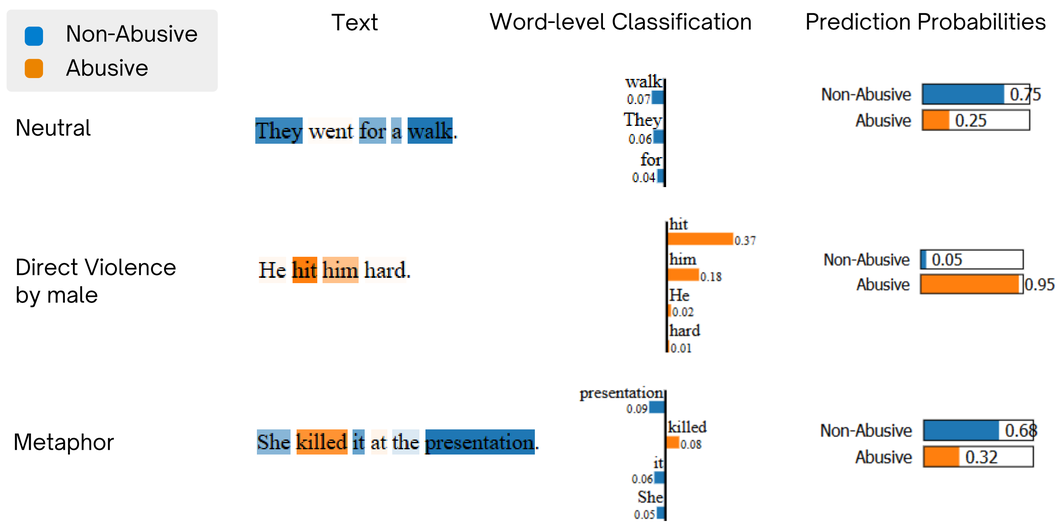
Top 10 terms for sexual violence:		
	Term	TF-IDF Ratio
255	mum	66.233
726	power	57.348
101	cat	51.667
52	non	50.593
452	chose	48.925
234	ruin	46.710
419	unhappy	46.435
608	age	45.678
378	positive	45.674
233	restraining	45.655

Top 10 terms for aggressive behavior:		
	Term	TF-IDF Ratio
503	mum	67.662
417	cat	56.882
764	death	48.482
746	spouse	47.958
160	proof	47.142
306	black	46.345
695	judge	46.308
663	neck	46.118
241	power	45.520
434	legal	45.429

Model Explanation

LIME Exploration



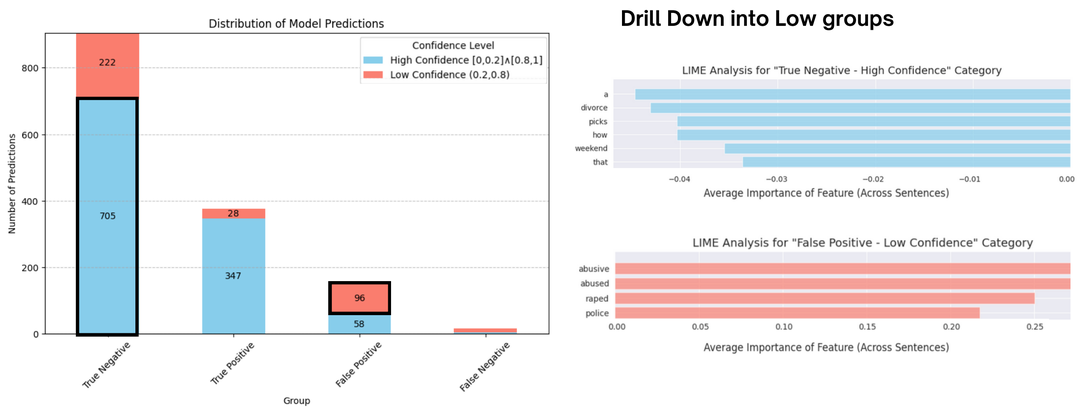


Confusing case:

I left an relationship I was in since I was 15 years old and this made me cry my eyes out I'm still in contact with my ex and I'm seriously considering cutting all contact even tho I'm extremely worried about what's gonna happen I dont want any family or friends getting involved as there all oblivious to the situation at hand I'm absolutely terrified someone is gonna get hurt or something bad is gonna happen but I can't keep on putting myself through this

Non-Abusive 0.49
Abusive 0.51

Prediction Explanation using Confidence Levels



Future Work

- Normalize **generic words** in the training set across labels.
- Add **generated data** to the training set to address the model's weaknesses in NLP (improve FP Low confident...)
- Compare the performance and conclusions to **different models.**
- Check the **Hebrew language.**
- Establish **practical actions** to help the victioms.