Domestic Violence Prediction

NLP Text Analysis of Reddit and Twitter Data

Elie Park
Mentor: Ben Bell

Reference Papers: Reddit paper[1], Twitter paper[2]

June 6, 2022

[1] Schrading et al. (2015). An analysis of Domestic Abuse Discourse on Reddit. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.* https://aclanthology.org/D15-1309.pdf

[2] Schrading et al. (2015). #WhyIStayed, #WhyILeft: Microblogging to Make Sense of Domestic Abuse. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics.* https://aclanthology.org/N15-1139.pdf

SUMMARY

In 2014, camera footage that showed American football player Ray Rice physically assaulting his then-fiancée, Janay Palmer was released to the public. They were subsequently married a month after the release. Rice apologized to his fans, his team, and his team's executives, but excluded his wife from the apology, while Palmer apologized for her role in the incident and focused more on Rice's professional career rather than his accountability. Social media users expressed outrage towards the NFL for not taking appropriate action and towards Palmer for staying in the relationship.

Later that same year, writer and domestic abuse survivor Beverly Gooden posted a series of tweets tagged #WhyIStayed to discuss the dynamic of her past abusive relationship and why it was so difficult to leave. The hashtag started to trend the same day and was used more than 100,000 times in less than two days after its creation, and #WhyILeft hashtags were also created the same night, with survivors discussing how they escaped from domestic violence. Such massive discourse on the social platform is valuable because it can be used to analyze the narratives of people who are likely to get into or stay in abusive relationships.

Domestic violence accounts for 21% of all violent victimizations in the US (2003-2012). Intimate partner violence accounts for 15% of these victimizations, in comparison to violence by immediate family (4%) or other relatives (2%). The pandemic was especially challenging globally for people in abusive relationships, with the outbreak contributing to increases of 18% police reports in San Antonia, 22% in Portland, Oregon; and 10% in New York City[3]. Globally there were upsurges of up to 300% incidents in Hubei, China; 25% in Argentina, 30% in Cyprus, 33% in Singapore, and 50% in Brazil[4].

In Canada, 44% of women and 36% of men who had ever been in an intimate partner relationship reported experiencing some kind of psychological, physical, or sexual abuse in the relationship[5]. In 2020, Canada's Assaulted Women's Helpline received 71,563 calls between April 1 and September 31. Compared to the 36,362 calls received in the same period, the number almost doubled during the pandemic. Call volumes also spiked immediately following the nation's first lockdown. Between March and June 2020, data from 17 police forces across Canada also showed that calls related to domestic violence rose by 12 per cent compared to the same four months in the previous year[6]. With quarantine regulations in place, it may have been more difficult for people to leave abusive environments. The nature of the virus also contributed to unique challenges in support shelters for these victims.

The pandemic also contributed to the increase in demand for mental health services. A recent survey by Mental Health America found that 54% of 11- to 17-year-olds reported frequent suicidal thoughts or self-harm in the two weeks before returning back to school after a year and a half of remote learning. This is the highest

---

[3] Boserup, B., McKenney, M., Elkbuli, A. (2020). Alarming Trends in US Domestic Violence During the COVID-19 Pandemic. *American Journal off Emergency Medicine. Vol. 38, Issue 12 2753-2755.* https://www.ajemjournal.com/article/S0735-6757(20)30307-7/fulltext

[4] Kluger, J. (2021). Domestic Violence is a Pandemic Within the COVID-19 Pandemic. *Times Magazine.* https://time.com/5928539/domestic-violence-covid-19/

[5] Cotter, A. (2021). Intimate Partner Violence in Canada, 2018: An Overview. *Canadian Centre for Justice and Community Safety Statistics.* https://www150.statcan.gc.ca/n1/pub/85-002-x/2021001/article/00003-eng.htm

[6] Thompson, N. (2021). Domestic Violence Reports Continue to Rise Due to COVID-19 Pandemic. *The Canadian Press.* https://www.cp24.com/news/domestic-violence-reports-continue-to-rise-due-to-covid-19-pandemic-1.5309133?cache=almzxqnumb%3FclipId%3D68596

rate since it began screening in 2014[7]. With reduced access to traditional forms of mental health resources, people aged 18-25 have turned to social media (51.4%) and mental health apps (32.6%) for support[8]. This method of reaching out for help may especially be essential for people who can't make a phone call due to an abusive partner being present during quarantine. Getting emotional support can be beneficial not just for the victim, but also for the abuser as intimate partner violence and mental health are closely connected.

7 Cups is one online platform where members can anonymously talk to a volunteer for free, or receive counseling with licensed therapists with a paid subscription. Volunteers can go through the Active Listening training program, where they can remotely practice compassionate and non-judgemental listening. However, if they encounter a situation which they are not licensed to deal with, such as suicide, they are encouraged to refer the member to a registered therapist. As such, because most volunteers are inexperienced, they might not be equipped to assess the risk levels of domestic abuse. We can use natural language processing to aid volunteers in understanding the narratives of people who stay vs. leave abusive relationships and providing the proper resources and emotional support for these victims.

Models

Three classifier models were tested: Naive Bayes Classifier, Random Forest Classifier, and Logistic Regression. The first case takes posts and comments tagged non-abuse or abuse from Reddit in order to identify if a dialogue contains abusive material. The second case takes the aforementioned tweets hash-tagged #WhyIStayed or #WhyILeft in order to determine if there is a consistent pattern within each narrative. The top two models, Naive Bayes and Logistic Regression, can be used in conjunction as an indicator of flagging users who potentially require professional help beyond the capacity of volunteers and chatbots.

Data

The Reddit data were downloaded from http://www.nicschrading.com/data/ as a shelved set of Reddit data, where abuse labels were allocated to submissions under subreddits: r/AbuseInterrupted, r/DomesticViolence, and r/SurvivorsofAbuse. The non-abuse labels were allocated to r/CasualConversation, r/Advice, r/Anxiety, and r/Anger. The data shelved under "submissions" and corresponding list of "comments" from Reddit were joined together into an even set of 552 abuse and 552 non-abuse posts which had at least one comment per post. This data did not contain any emojis.

The Twitter data was pulled from the Twitter API via the code on the same website using Twython, a personal Twitter API account and a json file. The dataset containing 30377 tweets had automated gold standard labels. Often, these tweets contained both hashtags, in which case they were split automatically with regexes. The data was cleaned independently, resulting in a final dataset of 19552 tweets, with 13731 #WhyIStayed and 5821 #WhyILeft tweets used for modeling. Emojis were replaced with the corresponding textual meaning using the demoji module. Mentions, hashtags, and redundant retweets were deleted. The original datasets were used by the author Nicholas Schrading in his original analyses of domestic abuse - his papers are linked on the cover of this pdf.

An emotions sensor dataset was downloaded from Kaggle at https://www.kaggle.com/datasets/iwilldoit/emotions-sensor-data-set as a .csv file for sentiment analysis of both reddit posts and tweets. It contains 1104 widely used words and their corresponding scores of 7 emotions: disgust, surprise, neutral, anger, sad, happy, and fear. The mean scores of each post and tweet were calculated using the scores for each token. For more accurate results, the complete dataset containing over 20,000 words should be used.

---

[7] Oliver, M. (2021). Pandemic Intensifies Growing Mental Health Crisis Among Teens. *CBS News.* https://www.cbsnews.com/news/mental-health-teens-covid-19-pandemic/

[8] Pretorius, C. & Coyle, D. (2021). Young People's Use of Digital Tools to Support Their Mental Health During Covid-19 Restrictions. *Frontiers in Digital Health.* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8671300/

REDDIT SUBMISSIONS: EDA

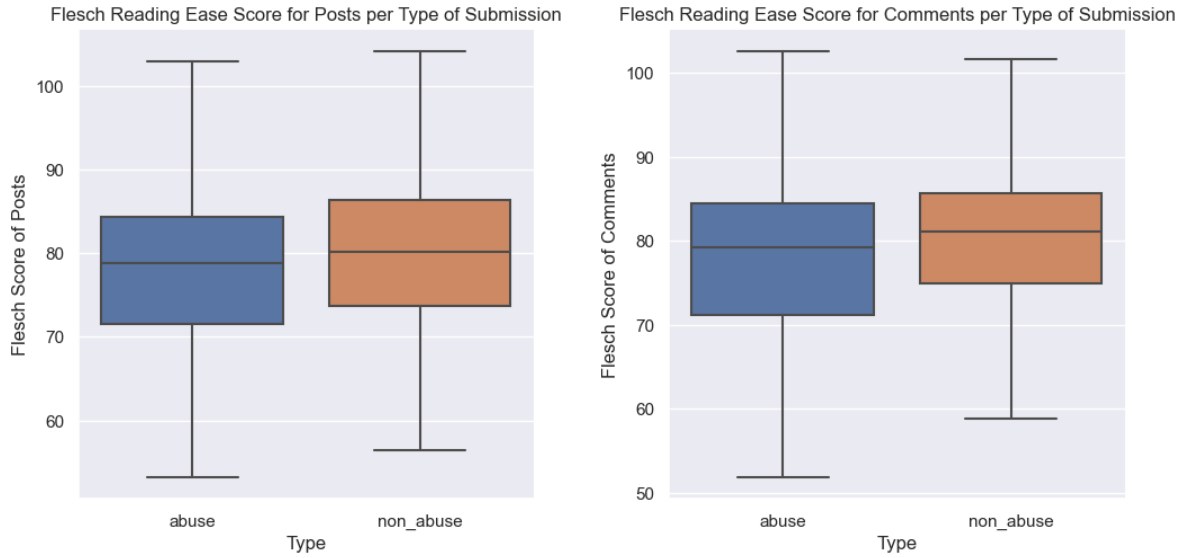


Figure 1. Distributions of Flesch Reading Scores of Reddit Posts and Comments

The reading scores were in the 70-85 range for abuse posts and comments and in the 73-86 range for non-abuse posts and comments. T-tests confirmed that the means of flesch scores in non-abuse posts and comments were higher ($p << 0.001$). The percentage of capital letters in non-abuse posts and comments were higher ($p < 0.01$).

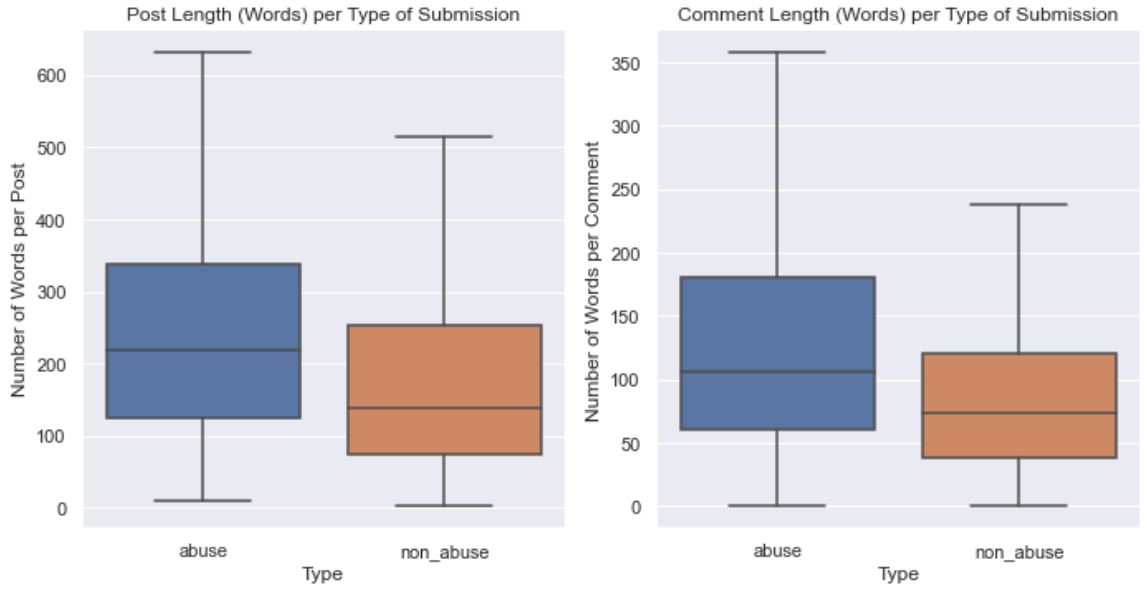


Figure 2. Distributions of Number of Words per Post and Comment

Posts and comments related to abuse were generally longer, containing more characters and words ($p << 0.001$). Comments under abuse posts averaged 54.8 words, while comments under non-abuse posts consisted of an average of 44.2 words. Average word length in posts were also higher for abuse posts and comments ($p << 0.001$).

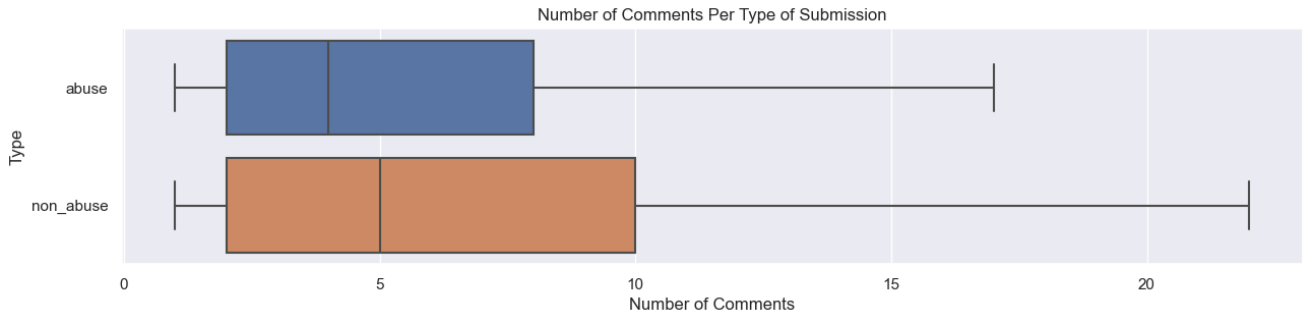


Figure 3. Number of Comments per Post

All posts in this dataset had at least 1 comment, and 50% of all posts contain 5 or less comments. The most commented post had 207 comments. These outliers (posts with many comments) were labeled non-abuse and contain questions directed towards other Redditors while sharing their own experiences. Posts about abuse tend to be more wordy (Figure 2), while having significantly less comments (Figure 3) ($p<<0.001$).

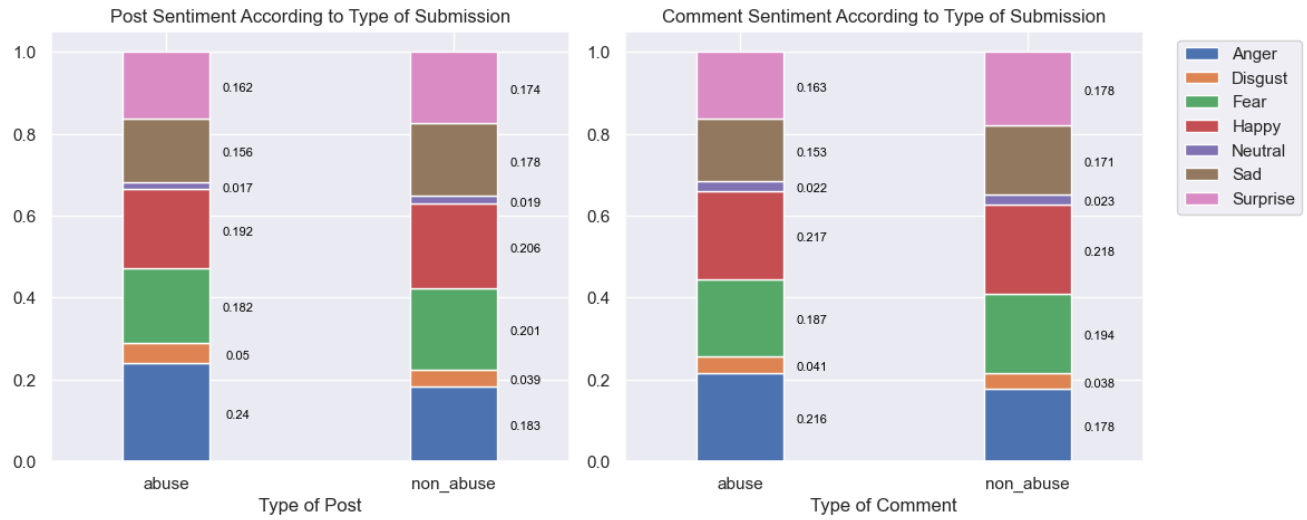


Figure 4. Relative Frequencies of Top Sentiments per Post and Comment

Sentiment scores were calculated for each post and comment with the emotions sensor dataset from Kaggle, using the mean scores of the emotions corresponding to each post and comment. Figure 4 shows the percent stacked bar charts of posts and comments using these scores. Out of the seven emotions, abuse posts scored the highest in anger as expected. Non-abuse posts scored fairly high in happiness and fear compared to other emotions because the author of the dataset purposefully included casual conversation, advice, anxiety, and anger posts as control variables. Abuse comments scored high in anger but also in happiness. Most happy words in comments were also positive words of comfort and advice, such as *love*, *happy*, and *protect*, but they were also used to express skepticism in being able to be happy.

REDDIT SUBMISSIONS: MODELING

Model	Best min_df for CountVectorizer	Best Parameters	ROC-AUC Score
Naive Bayes Classifier	**1**	**{'alpha': 1}**	**0.987**
Random Forest Classifier	14	{'criterion': 'entropy', 'max_depth': None, 'max_features': 'log2', 'n_estimators: 700}	0.979
Logistic Regression	1	{'C': 0.1}	0.973

Table 1. Comparison of Three Classifiers

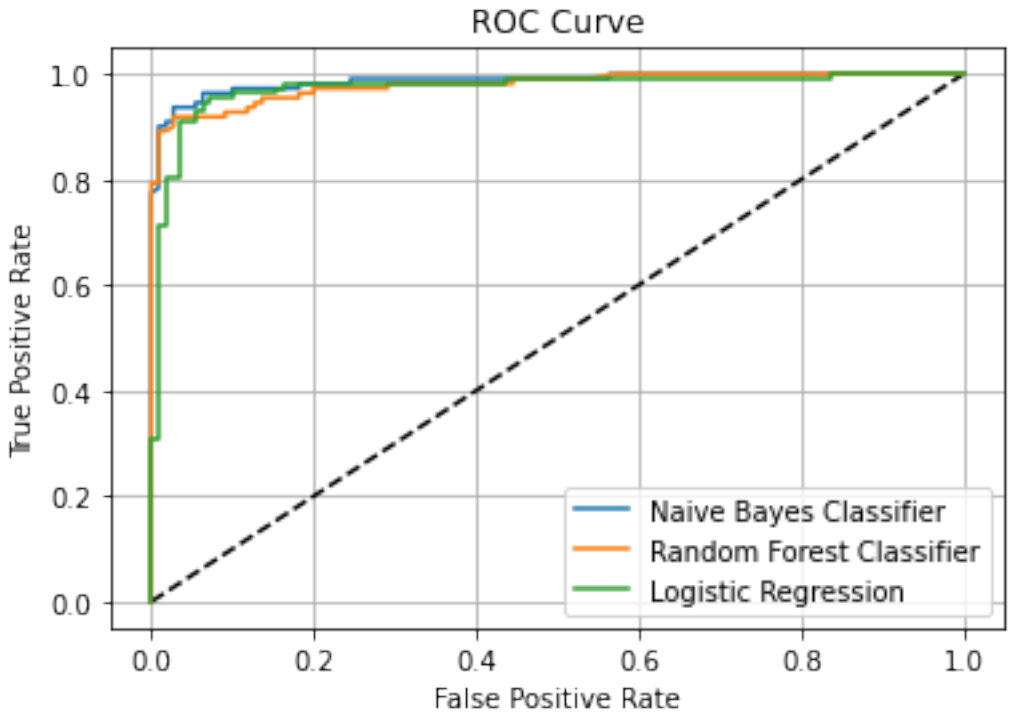


Figure 5. ROC Curves of Three Classifiers

The Naive Bayes Classifier proved to be the most effective model in predicting abuse vs. non-abuse out of the three models. The 'min_df' parameter for CountVectorizer which is used to convert the text into a vector of token counts were tested with all three models within a range of 0 to 20 for posts, comments, and posts and comments combined. Combining both posts and comments resulted in the highest test accuracy for all models. The three models had high ROC AUC scores as shown in Table 1 and the best parameters were found using gridsearchCV. This corresponds to the close proximity of the ROC curve to the top-left corner in Figure 5.

REDDIT SUBMISSIONS: FEATURE IMPORTANCES

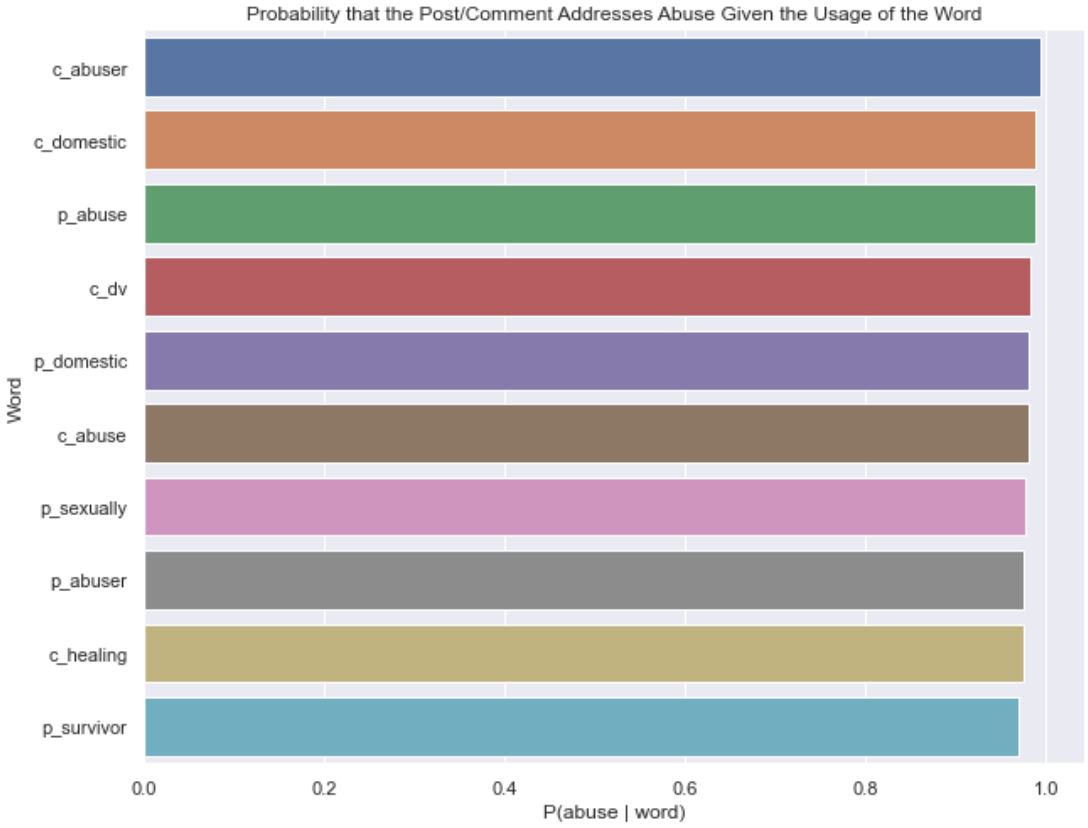


Figure 6. Most Predictive Words of Abuse Posts and Comments

Using the Naive Bayes Classifier, the top 10 most predictive words for predicting abuse were outputted. Tokens from posts are denoted with the prefix 'p_' and tokens from comments were denoted with 'c_'. The results are not very surprising, except 'c_healing' which contained words of encouragement in comments under abuse posts.

> "Depending on where you live, there may be a law in place that releases you of a lease if you've experienced **DV**, reporting or not. Talk to your local **DV** agency. They'll be able to give you local resources and help inform you of your rights and options." (Comment)

> "Being abused gives us a different perspective on life. It is not necessarily something that will ever go away because it almost becomes apart of you but things will become better as time goes on. When you focus on your **healing** you will begin to feel happy and trustful of the world once again." (Comment)

> "When the app is activated it will start to record audio from the phones microphone and send out a text message to an emergency contact (or contacts) with a customisable message along with the location of the **survivor**" (Post)

In future analyses, we may decide to exclude some comments that sound more like advice than the narratives of the victim. This could be accomplished with a crowd-sourced annotation study.

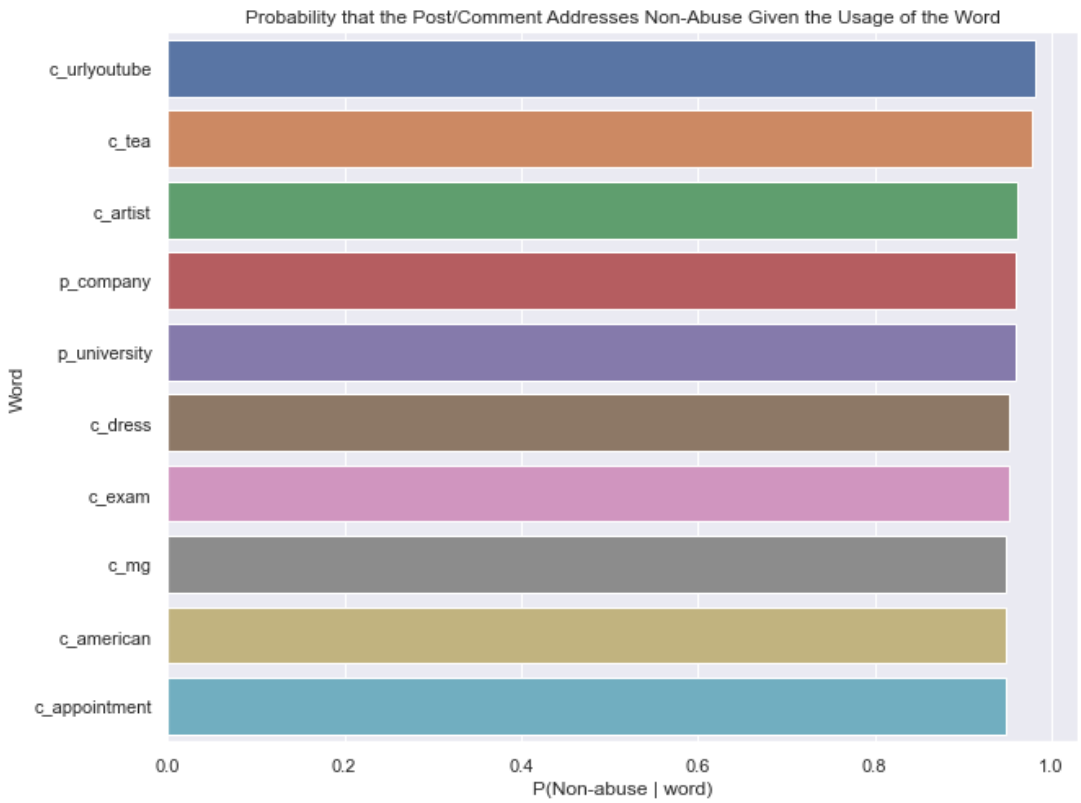


Figure 7. Most Predictive Words of Non-Abuse Posts and Comments

Tokens from comments were more useful in predicting non-abuse, because although there were less number of words overall per each comment than per post, there were more comments under non-abuse posts (Figure 3). The most predictive token was a url link to YouTube due to the fact that some posts containing the most comments were threads related to music:

> “In this thread: songs that invoke massive feels. Any genre, any type of feeling. Super simple but good lord, I am listening to a massive feels song.” (70 Comments)
>
> “Anyone want to talk about rap music? Who's your favorite rapper? Least favorite? Why?” (207 Comments)

The comments under these posts were as such:

> “For a real song that always gives me feels I choose [this](https://www.youtube.com/watch?v=9pkLDEEs20U).”
>
> "I\'m really into [shad](https://www.youtube.com/watch?v=v5oJMXqxsOQ&list=PLE7D70099B3F7CBF5) right now. I used to like busta rhymes before I hit puberty but not anymore. He goes way to fast for anything to be coherent on the first listen."

Tea was used r/CasualConversation to discuss what type of tea people liked:

> "Don't have any pictures to post but I was drinking some **tea** and described the flavor as apricot and my friend honestly didn't even know what an apricot was." (Comment)

Many posts with *mg, and appointment* were related to anxiety:

> "Zoloft was horrible for me for the first two weeks. I was extremely anxious and unable to sleep. It was just awful. In my case, it was because I had to increase the dosage quickly, all the way up to 150 **mg** in just over a week. I got prescribed trazodone for sleep, which helped somewhat, but really I just had to wait for the side effects to go away. I'm glad I did, because I feel much better now." (Comment)

Company, university, *exam*, and *dress* were used to ask for or give advice:

> "I had a similar experience the other day. The way uni **exams** work in Italy is that there are several sessions and you get to pick when you want to take your exam. During the **exam** you can "retire" (translated literally), which makes you eligible to retake that **exam** at another time. " (Comment)

Other words were random in context:

> "Hey, look at it this way: it's your first draft! :) It's bound to be riddled with mistakes and whatnot. Don't worry about how bad it is; once you receive constructive criticism on your work, you'll be able to improve on your weaknesses and focus on your strengths. Writers, just like any **artist** or singer or doctor out there, get better the more they practice." (Comment)

Non-abuse posts and comments seem to be indicative of the choice of control submissions used in the dataset.

REDDIT SUBMISSIONS: THRESHOLDING

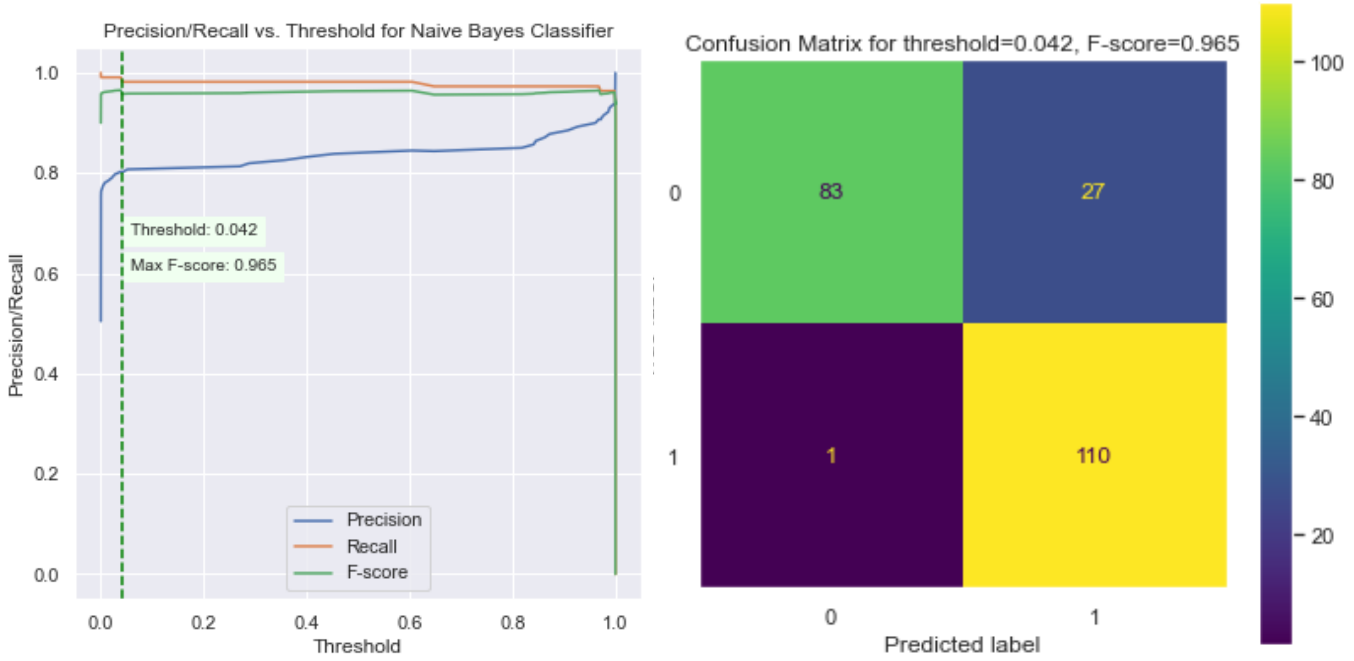


Figure 8. Precision/Recall vs. Threshold and the Confusion Matrix for Selected Threshold

	Precision	Recall	F1-Score	Support
0	0.99	0.75	0.86	110
1	0.80	0.99	0.89	111
Accuracy			0.87	221
Macro Avg	0.90	0.87	0.87	221
Weighted Avg	0.90	0.87	0.87	221

Table 2. Classification Report

Non-abuse posts and comments have been labelled 0 and abuse posts have been labelled 1. The Multinomial Naive Bayes Classifier mislabeled non-abuse posts as abuse more frequently than it mislabeled abuse posts. For our case of identifying narratives of abuse, we deem this model important when identifying abuse, rather than non-abuse, as discussed in Figure 7. Therefore, we aim to minimize false negatives - abuse posts and comments falsely mistaken for non-abuse posts and comments, by placing more importance on recall and less on precision. Precision stays above 0.8 and recall stays above 0.95 for most threshold values. A beta of 2.8 was used to calculate the threshold of 0.042 corresponding to the highest F-beta score of 0.965. We were able to achieve a precision of 0.8 and a recall of 0.99.

TWITTER DATA: EDA

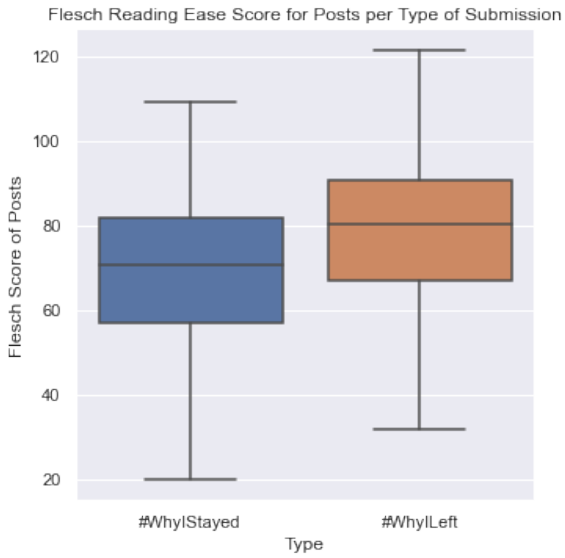


Figure 9. Distributions of Flesch Reading Scores of Tweets

#WhyIStayed reading scores were mostly in the 57-70 range and #WhyIStayed reading scores were mostly in the 67-90 range. T-tests confirmed that the means of flesch scores in #WhyILeft tweets were significantly higher ($p << 0.001$. This result may be attributed to that #WhyIStayed tweets contained more urls and mentions than #WhyILeft tweets, which was confirmed with a chi squares test ($p << 0.001$). #WhyIStayed tweets also consisted of more capital letters ($p << 0.001$). There was no statistically significant difference between the counts of emojis in #WhyIStayed and #WhyILeft tweets.

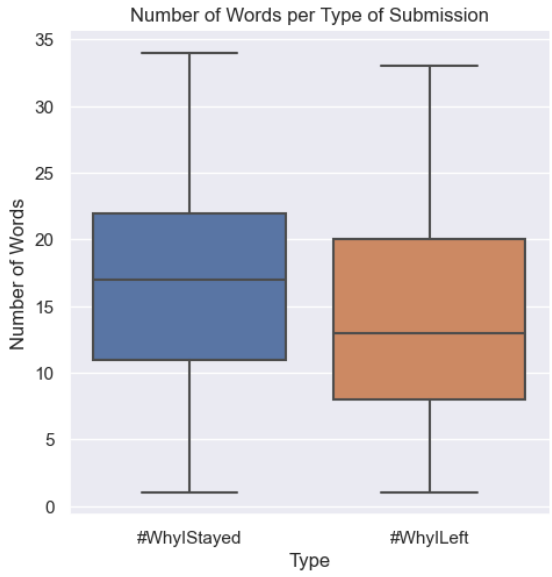


Figure 10. Distributions of Number of Words per Tweet

#WhyIStayed tweets contained more words and tokens than #WhyILeft tweets.

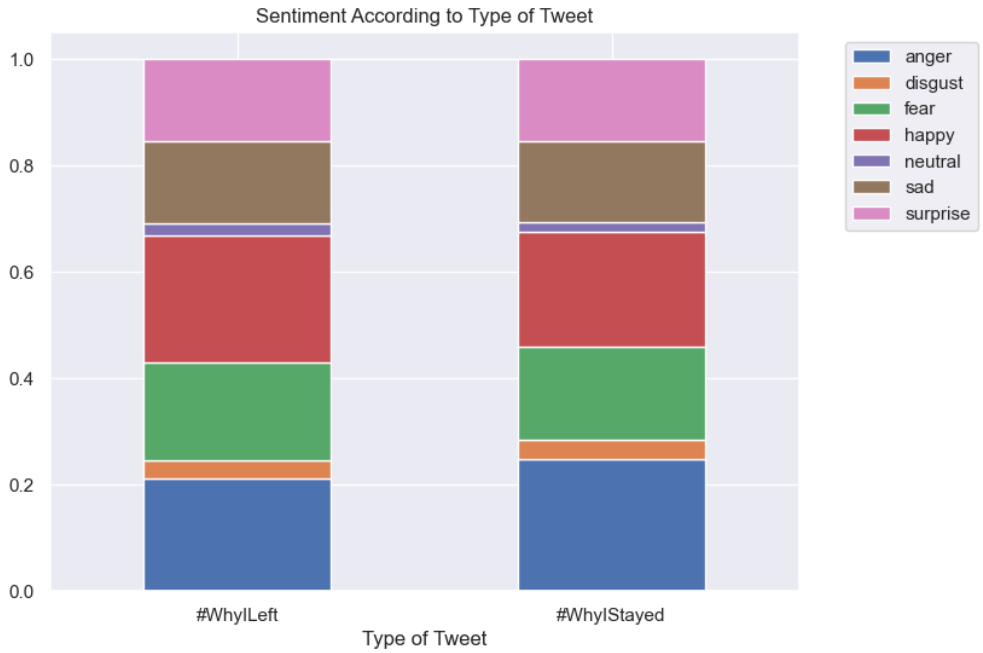


Figure 11.
Relative Frequencies of Top Sentiments per Tweet

#WhyILeft tweets had a higher percentage of the happy sentiment. *Love*, the most common word in both labels of "happy" tweets, is directed towards the self in #WhyILeft, but in contrast, it is directed towards the abuser in #WhyIStayed. *Better* is the second common word and is used in a hopeful manner towards the future in #WhyILeft, but it is often used as a personal adjective in #WhyIStayed. *Daughter* was a factor in why people left abusive relationships, but *marriage* was also an indication of why people stay.

#WhyIStayed tweets had a higher percentage of the angry sentiment. They often contained recollections of why victims stayed, years after the abuse. *Abuse* and *leave* were the most common words in both types of angry tweets. #WhyILeft tweets demonstrate a change in perspective in viewing abuse, while #WhyIStayed tweets denote a lack of inner awareness and external validation. The word *wrong* in #WhyILeft tweets was used as a direct identifier of a situation, but in #WhyIStayed tweets, it suggests an absence of this identification.

TWITTER DATA: MODELING

Model	Best min_df for CountVectorizer	Best Parameters	ROC-AUC Score
Naive Bayes Classifier	1	{'alpha': 1}	0.824
Random Forest Classifier	9	{'criterion': 'entropy', 'max_depth': 30, 'max_features': 'log2', 'n_estimators: 1000}	0.821
Logistic Regression	**1**	**{'C': 1}**	**0.829**

Table 3. Comparison of Three Classifiers

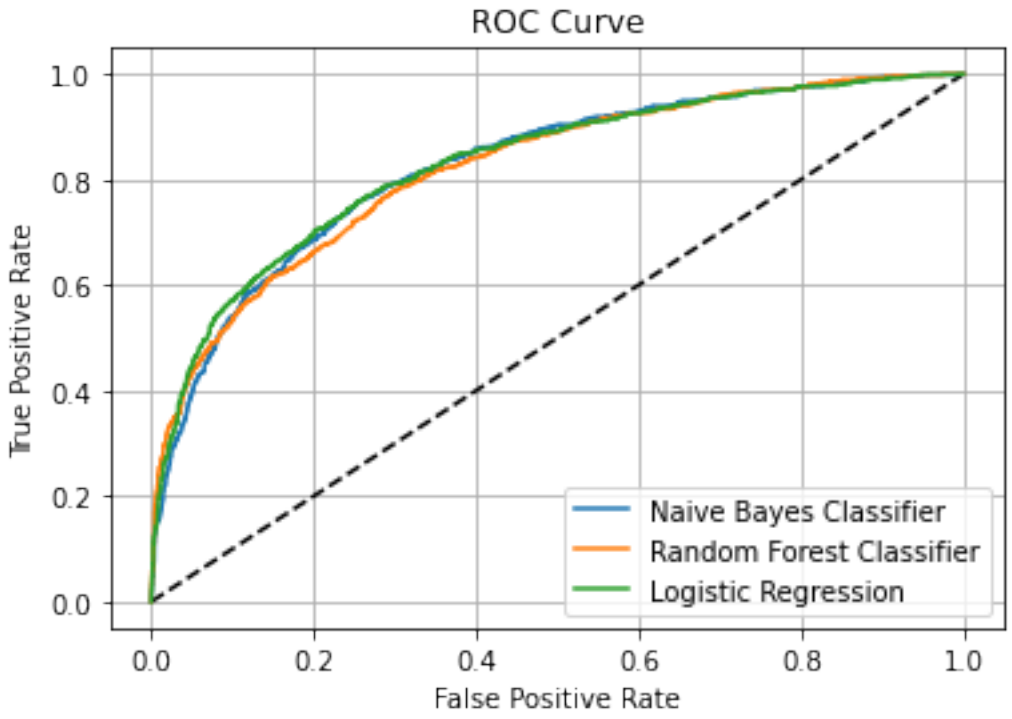
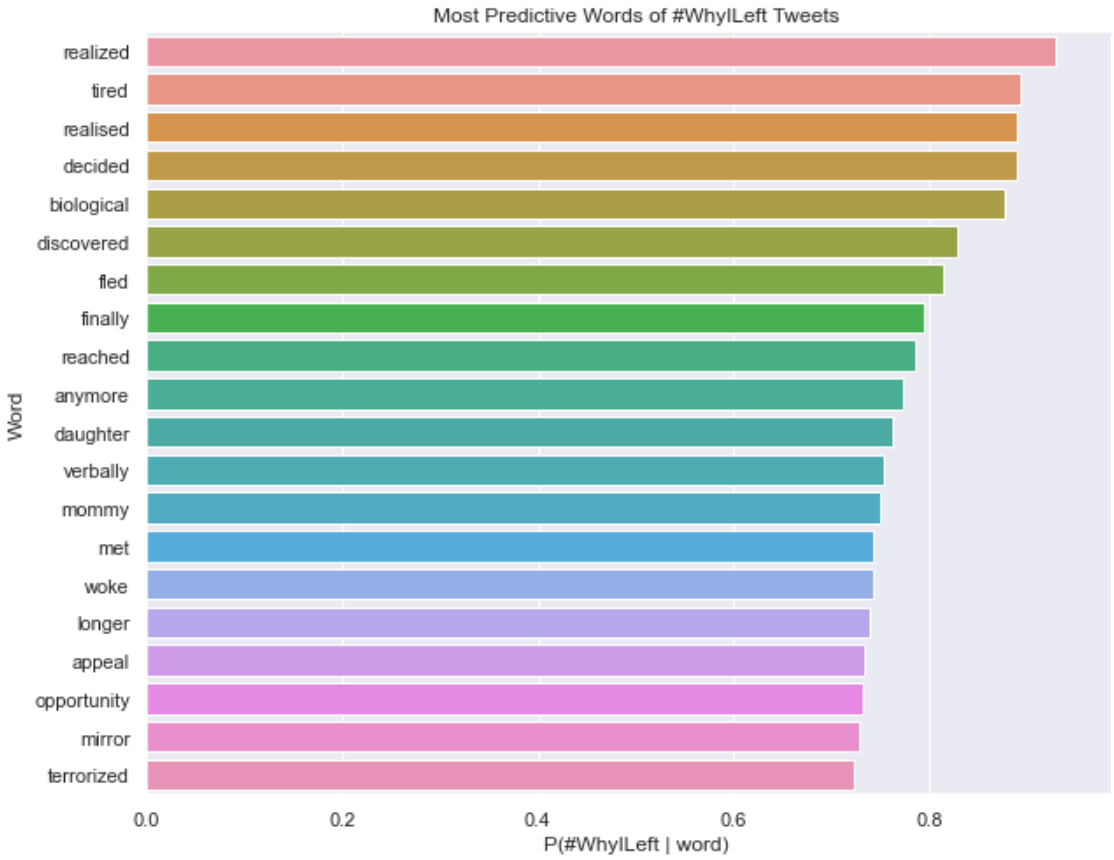


Figure 12. ROC Curves of Three Classifiers

A total of tokens extracted from the Reddit data were used in modeling. Logistic Regression proved to be the most effective model in predicting #WhyILeft vs. #WhyIStayed out of the three models. The 'min_df' parameter for CountVectorizer were tested with all three models within a range of 0 to 20. The three models had very close ROC AUC scores as shown in Table 1 and the best parameters were found using gridsearchCV, similarly to the Reddit dataset. The difference between using Logistic Regression and Naive Bayes is that Naive Bayes assumes all the features to be conditionally independent, and Logistic Regression works better than Naive Bayes even when some of the variables are correlated. The highest score in Logistic Regression for this particular dataset is surprising in comparison to the highest score of the Naive Bayes for the smaller Reddit dataset, because Naive Bayes generally does better with less variables and data.[9]

[9] Schapire R. & Blei D. (2007). Logistic Regression and Naive Bayes (Rob). *Princeton University.* https://www.cs.princeton.edu/courses/archive/spr07/cos424/scribe_notes/0410.pdf

TWITTER DATA: FEATURE IMPORTANCES



SFigure 13. Most Predictive Words of #WhyILeft Tweets

Using Logistic Regression, the top 20 most predictive words for predicting the label #WhyILeft were outputted. *Realized*, *tired*, *decided*, *discovered*, *finally*, and *anymore,* describe both gradual and sudden moments of breaking out of learned helplessness through self-empowerment:

> “"@ravenmundy: #whyileft Because I **realized** someone had to actually love me, even if it was only myself." Wow! #truth”

Finally, mirror, anymore, woke, and *terrorized* are contained in often graphic tweets describing physical and emotional abuse:

> “#whyileft My mental and physical health was bottoming out. I ended up in the hospital. I couldn't do it **anymore** even if I loved him."

While most of these tweets described intimate partner violence, some tweets used *biological* to particularize violence by blood-related family members. *Reached, met, opportunity, and fled are used to describe how outside influences such as family, friends, and new relationships intervened with the situation. Some of these influences were more abstract (eg. Dreams, empathy).*

> “#whyileft An old friend **reached** out, showed me that I was not alone, and helped me find a way out.”

Some of the outside intervention was defined by the presence of children.

> “#whyileft my son refusing to go to school cos "daddy hits **mommy**" was the last straw. I finally saw it through his eyes. Near death. Twice”

Appeal was used only in two posts tagged #WhyILeft and no posts tagged #WhyIStayed.

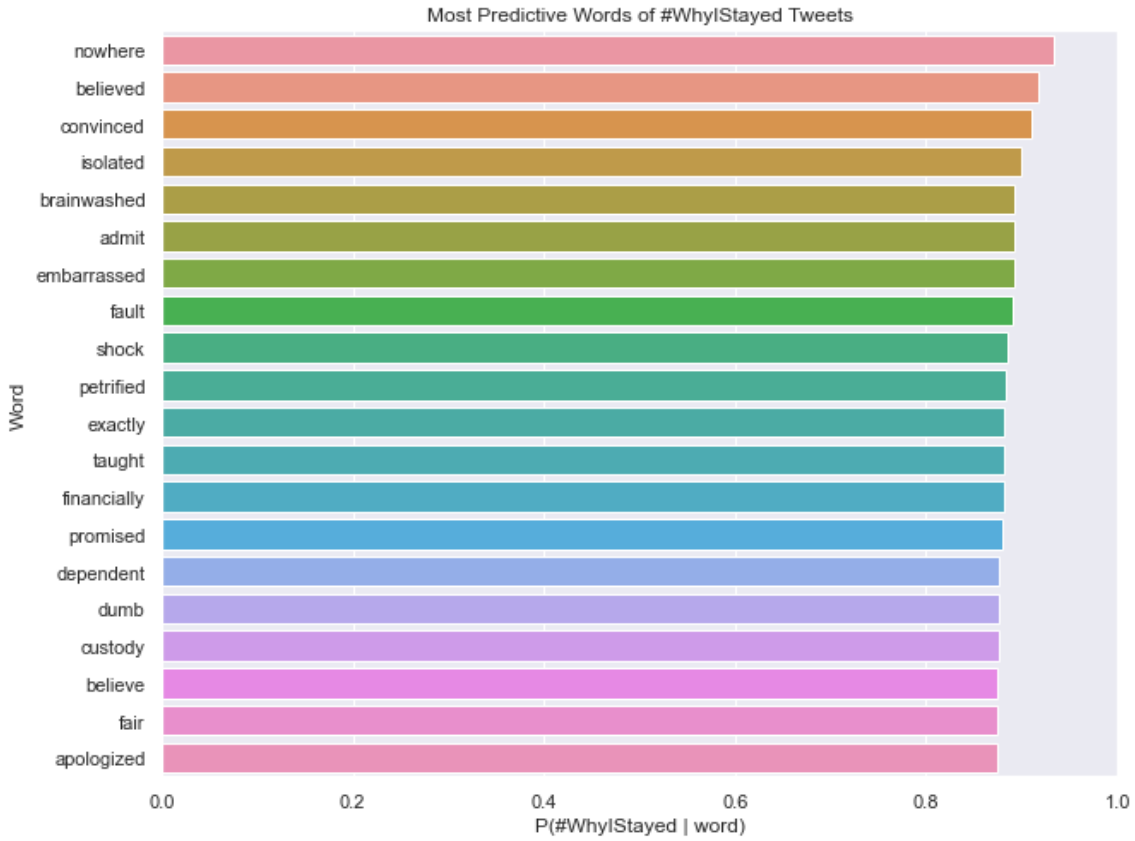


Figure 14. Most Predictive Words of #WhyIStayed Tweets

The reasons why people stay in abusive relationships can be categorized into many narratives. The top words *nowhere*, *believed, convinced, isolated, fault, exactly, dependent,* and *dumb* were used to convey the erosion of self-worth through repeated exposure to name-calling and criticism. Insecure attachment to the abusive partner is described in these tweets.

> “#WhyIStayed I felt **dependent** on him. After you are told and treated like you are worthless repetitively enough you start to **believe** it.”

Believed, convinced, and *dumb* were features illustrating lack of social/outside validation. Some of these tweets described victims turning a blind eye to emotional and verbal abuse because there was no physical abuse. Other tweets recount situations where people outside of the relationship had a certain disregard for what was happening. Often this disregard was internal and rooted in a history of family violence.

> “#WhyIStayed My parents hit me & said they loved me. So when he hit me & said he loved me I **believed** him.”

Victims often experienced feelings of shame before making their choice to stay.

> “The reason #WhyIStayed after he moved his girlfriend in was because I really had no place to go and I was too **embarrassed** to tell anyone.”

> “Because I was a strong intelligent woman and I didn't want people to know I was **dumb** enough to let it happen to me. #WhyIStayed"”

Nowhere, financially, brainwashed, isolated, and *dependent* describe situations where the victim did not have control of his or her finances, and felt socially isolated from loved ones. Many tweets relating to finances recalled experiencing economic abuse, where one partner controls the other partner’s ability to acquire, use, and maintain economic resources to prevent victims from leaving the abuser by lack of financial means.

> “Being **isolated** enough to hate my friends, family, and self, while **brainwashed** into thinking that was how all relationships are #WhyIStayed”

> “He also made me quit working, made me completely **dependent** on him, &; I realized all too late he had cleaned me out **financially.** #WhyIStayed”

Some of the tweets recalled abusive partners using threats against his or her own life, or towards the author and author’s loved ones.

> “#WhyIStayed because he threatened to kidnap the children or destroy me in a nasty **custody** battle if i ever left.

The presence of children affected these victims differently than those who left the relationships:

> “#WhyIStayed I felt **financially** trapped even though I knew I could support myself. It was different with a child with special needs though.”

> “#WhyIStayed cuz i **believed** I needed to give our children a whole family. No matter how miserable I was....I was wrong, left 2 years ago”

The abusive partner often apologized, but did not change their behaviours. *Apologized, promised, believed,* and *convinced* tweets demonstrated forgiveness, hope, and empathy towards the abuser.

> “#WhyIStayed After he was violent, he **apologized** and was extremely apologetic and loving, until the next blowup.”

Although there were no words related to the church in the feature importances, a similar theme in the keywords *taught* and *fault* were regarding religious views of marriage, as many religions consider divorce to be sinful.

> “#WhyIStayed my pastor told me to pray more, fast more, submit more. I was told by my church it was my f**ault.**”

Additional tweets are shown in the Appendix.

TWITTER DATA: THRESHOLDING

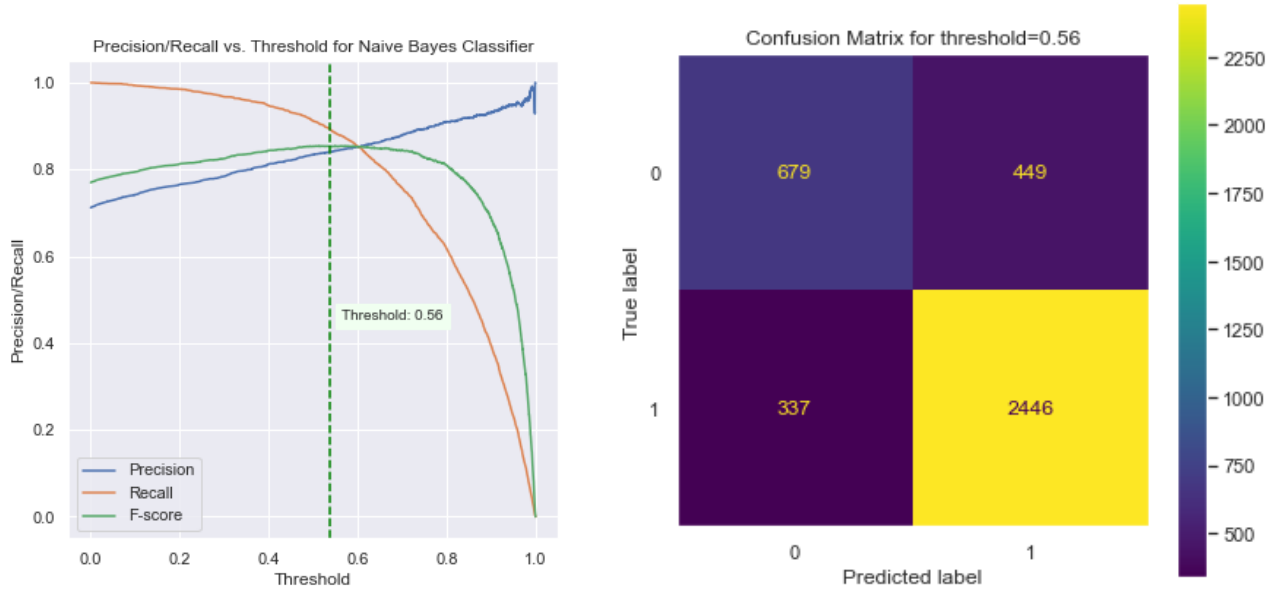


Figure 8. Precision/Recall vs. Threshold and the Confusion Matrix for Selected Threshold

	Precision	Recall	F1-Score	Support
0	0.67	0.6	0.63	1128
1	0.84	0.88	0.86	2783
Accuracy			0.8	3911
Macro Avg	0.76	0.74	0.75	3911
Weighted Avg	0.79	0.8	0.8	3911

Table 2. Classification Report

#WhyILeft tweets have been labelled 0 and #WhyIStayed tweets have been labelled 1. For our case where we employ this Logistic Regression model alongside the Naive Bayes model (Reddit), we are more interested in warning users of narratives of #WhyIStayed instead of #WhyILeft. Therefore, we put more emphasis on recall than precision. Figure 8 shows the recall increasing but the precision decreasing even more as we decrease the threshold. Picking a threshold of 0.56 allows us to attain a recall of 0.88 with a precision of 0.84 without lowering the recall of #WhyILeft tweets under 0.6. With a lower chosen threshold, we are able to increase recall, but the majority of #WhyILeft tweets are also incorrectly labelled #WhyIStayed.

CONCLUSION

The Twitter analysis allowed for a deeper understanding of the narratives of people who stayed vs. left abusive situations. Despite the severity of domestic violence, victims were not able to leave because of dependencies, lack of validation/boundaries, fear of retaliation, custody issues, forgiveness/hope, and religion. The psychological, financial, emotional, and physical dependencies were often caused by their partners seeking a sense of control over the victims, and having grown up around abusive situations worsened the case by influencing their perception of what was considered "normal" in current relationships. These symptoms were mostly congruent with *battered person syndrome*, which is a psychological condition that can develop when people experience domestic abuse, exhibiting four characteristics: they believe violence is their fault, they can't place the blame for the violence on anyone else, they fear for their lives and their children's lives, and they believe their abuser is everywhere and sees everything they do.[10]

#WhyIStayed was also used to criticize the movement of putting focus onto victims, rather than the abusers, although that was not the intent behind its creation. Some people were on the other side of the argument and tweeted that it is indeed important to understand the victim's perspectives in order to ease the victim-blaming. #WhyILeft tweets mostly described moments of self empowerment and an awakening out of unhealthy perceptions of relationships and boundaries.

Although the Reddit analysis was not very revelational, the Multinomial Naive Bayes model for the Reddit data was helpful in identifying abuse cases with a precision of 0.8 and recall of 0.99. The non-abuse section of this dataset does reflect the types of conversations which may occur on an online counselling platform. This model can be used to determine if a conversation pertains to domestic abuse, alongside the Logistic Regression model for the Twitter data to identify the probability that someone could stay in an abusive situation with a precision of 0.84 and recall of 0.88. Despite the possibility to achieve a higher recall with a high precision, which allows the user to be alerted in almost all cases where victims show signs of staying within an abusive relationship, the corresponding lower threshold was not chosen because of the high probability that the user may be alerted even when a victim talks about leaving. Therefore, these models should not be employed in an unsupervised, automated setting.

In future analyses, we may consider filtering the comments under Reddit abuse submissions to isolate the victim's replies from ones giving advice and encouraging words through a crowd-sourced annotation study. Additionally, using a balanced Twitter dataset with even #WhyIStayed and #WhyILeft tweets may aid in increasing recall without also increasing the amount of false positives in the Logistic Regression model. The improved models may be used in future implementations of mental health conversational AIs.

[10] Orenstein, B. W. (2014). Understanding Battered Woman Syndrome. *Everyday Health.* https://www.everydayhealth.com/news/understanding-battered-womens-syndrome/

APPENDIX

Warning: The following material contains graphic descriptions of suicide and violence which some readers may find disturbing.

#WHYILEFT

“#whyileft I left him because I had enough with the insults and the threats and the empty demonstrations of so called love. I was **tired.**”

“#whyileft because I woke up one morning and **decided** to take my power back. I needed to set an example for my kids abuse is not acceptable!!”

“I **discovered** one day that I was powerful and beautiful, that my friends loved me. #WhyILeft”

"I **woke** up on the floor 4 hours after he knocked me out. I believe he thought he killed me. I knew when he came back, he would. #whyileft”
“He broke my nose, totally **terrorized** me after and threatened to kill my parents.”

“I left my **biological** family because I WAS NEVER LOVED BY THEM, NEVER EVER, I WAS USED AND ABUSED AND EVEN RAPED BY THEM”

“#WhyILeft Because I saw my deceased father in a dream and **woke** up knowing he would never want me treated that way.”

“#WhyILeft My now-wife gave me the courage and **opportunity** to leave.”

“I left because I had the support of my mother, who immediately let me move back home after I **fled** once he fell asleep. #whyileft”

“I didn't want my **daughter** growing up thinking it was "normal" for your husband to abuse you **verbally** and physically #WhyILeft”

"when he kicked my **daughter** in the stomach and said he'd put a bullet through my head, i knew it wasn't getting better. #WhyILeft"

#WhyILeft I finally found the word No once I had kids – they made me stay **longer** at first, but seeing the wrong with their eyes opened mine.

“To give my son the fair **opportunity** to learn to respect women, mothers, and me. And because I love him. #WhyILeft http://t.co/t3LRhoNkVl"

#WHYISTAYED

“She **apologized,** told everyone in her life what she'd done, and I was still too afraid to leave bc I didn't feel worthy”

“Because he told me that no one would else would ever want me &; claimed that I belonged to him. I was scared, so I **believed** him. #WhyIStayed”

“I had been conditioned to think it was my **fault.** Had to figure out it wasn't me and I had no reason to stay. #WhyIStayed”

“#WhyIStayed Because I've been told all my life that a man's only worth is in how he supports and protects women, and I foolishly **believed** it”

“I stayed because I **convinced** myself verbal abuse wasn't that bad if it didn't turn physical &; he said no one else would want me. #WhyIStayed”

“His mom **convinced** me to stay. Her need for s/one to babysit her son was greater than my safety. Even after she knew he beat me. #WhyIStayed”

“Because my mother said "you were **dumb** enough to marry him, now you have to suffer the consequences" #WhyIStayed”

“#WhyIStayed I was too stubborn to **admit** that my family was right”

“he had everyone i trusted **convinced** i was nuts so that when i was finally ready to reveal/get help. i would have **nowhere** to go. #WhyIStayed”

“#WhyIStayed we worked together, lived together. she controlled me **financially**, emotionally. she manipulated me into thinking i owed her”

“#WhyIStayed Because he told me he would kill me and that no one would ever love me the way he did. I was scared, felt **isolated** from family.”

“#whyistayed because I was afraid he would kill himself and that it would be all my **fault.** He did, and I am not to blame.”

“#WhyIStayed because I was in **shock**, like a nightmare, then when I gained focus he threatened to kill me.”

“He said he would change. He **promised** it was the last time. I **believed** him. He lied. #WhyIStayed”

"family court doesn't believe u if u have no proof, so they take your kids away from you, or force you to agree to joint **custody** #WhyIStayed"

“@Katiebyeager: I thought my daughter deserved a **fair** shot at a "happy family" #WhyIStayed” same with my son”

“#WhyIStayed he always **apologized** with gifts and promised it would never happen again. (Or at least until the bruises healed)”

“#WhyIStayed Because the elders of our church told my mother and I that it was our **fault.**”

“My mom stayed because my dad controlled the money, **church** told her to obey & submit, he **convinced** her it was best for the kids #WhyIStayed”

“#WhyIStayed being a Christian had **taught** me i had no value without my wife and that it could be "that bad”