8-2018

SecondLook: A Prototype Mobile Phone Intervention for Digital Dating Abuse

Tania Roy
Clemson University, taniaroy89@gmail.com

SECONDLOOK: A PROTOTYPE MOBILE PHONE INTERVENTION FOR DIGITAL DATING ABUSE

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Human Centered Computing

by
Tania Roy
August 2018

Accepted by:
Dr. Larry F.Hodges, Committee Chair
Dr.Jerome McClendon
Dr. Bart Knijnenberg
Dr.Shaundra B. Daily
Dr.Sabarish Babu

Abstract

Digital dating abuse is a form of interpersonal violence carried out using text messages, emails, and social media sites. It has become a significant mental health crisis among the college-going population, nearly half (43%) of college women who are dating report experiencing violent and abusive dating behaviors. Existing technology and non-technology based intervention programs do not provide assistance at the onset of abuse.

The overall goal of this dissertation is to create a mobile phone application that consists of a detection tool that classifies abusive digital content exchanged between partners, an educational component that provides information about healthy relationships, and a list of nearby resources for users to locate help.

For the user-interface design of this application, we conducted a focus group study and incorporated the themes generated from the study to create our Android prototype. We used this prototype to conduct a usability study to evaluate the overall user-interface design and the effectiveness of the features we incorporated into the app. Due to the lack of a publicly available dataset that could be used to create training and testing sets for the classifiers to detect abusive vs non-abusive text messages in the context of digital dating abuse, we first created and validated a dataset of abusive text messages. This dissertation describes the dataset creation, validation process and the results of an evaluation of different classification and feature extraction techniques. The combination of linear support vector machine, unigram input and tf-idf feature extractor with an accuracy of 91.6% was the most balanced classifier, classifying abusive and non-abusive text messages equally well.

Finally, we conducted a user study to investigate different visualization paradigms that will assist users to trust the feedback regarding the possible abusive nature of their online communication. Three different visualization techniques were evaluated using survey questionnaires to understand

which one is the most effective in invoking user trust and encourages them to access resources for help.

Dedication

”Where the mind is without fear and the head is held high,

Where knowledge is free

Where the world has not been broken up into fragments

By narrow domestic walls

Where words come out from the depth of truth

Where tireless striving stretches its arms towards perfection

Where the clear stream of reason has not lost its way

Into the dreary desert sand of dead habit Where the mind is led forward by thee

Into ever-widening thought and action

Into that heaven of freedom, my Father ...,”

—Rabindra Nath Tagore, *Translation from Naibadya*

To all the courageous survivors of interpersonal violence. #MeToo

Acknowledgments

Thank you to the School of Computing, Clemson University that supported my doctoral program.

Thank you to my advisor Dr. Larry F. Hodges for guiding and supporting me through my journey first in the Cyber Innovations Lab and then as his Ph.D. student. It was a privilege and joy to learn from you. You allowed me to grow both as a person and researcher through these years. No matter how difficult the journey became, you never let me lose focus, and your encouragement helped me pushed the boundaries of my capability, and I could not have completed without this. You are the best adviser I could have ever imagined. You and Mrs. Hodges opened your home to me even before I became one of your students. I am incredibly blessed to be able to call you both family.

Thank you to Dr. Juan Gilbert for introducing me to the world of Human-Centered Computing and giving me an opportunity to join the Ph.D. program. I would like to thank Dr. Shaundra B. Daily for taking me on as her student in spite of me not having any prior research experience. You allowed me to explore different avenues until I narrowed down the domain I wanted to work in. Dr. Melva James, thank you for being a supportive mentor and teaching me the 'ropes.' I would like to thank all the staff and faculty members of the School of Computing, especially Ms. Christy Babb for always being a source of support and encouragement I would like to thank my dissertation committee members Dr. Sabarish Babu, Dr. Bart Kninjenberg and Dr. Jerome McClendon for their invaluable expertise and support. To Dr. Sekuo Remy, thank you for all your encouragement and help with machine learning. I would not have taken this road without your assistance. Dr. Melinda Weathers thank you for helping me shape my research domain, your vision for this project formed the building blocks of my dissertation. Dr. Pradip Srimani, thank you for being the person I can always go to and get unbiased advice. Other than being the first professor I met in Clemson, you were my sounding board when I needed to know the actual merit of specific situations, and you gave me the most honest perspectives. I am grateful to have you as a mentor. Dr. Christopher Ruth and Dr. Mouzan, you both helped me preserve my mind during the ebb and tides of my journey. Thank you for supporting me through this journey, I am grateful for your kindness and encouragement. All the members of the Virtual Environments Group (Matias, Zach, Himanshu, Dhaval, Ayush, and Alex) - thank you for all the times I knew I could count on each and every member. The time I spent with all of you in classes, doing projects, working on paper deadlines or sharing life highs and lows have been invaluable, and I will treasure every moment.

To each and every member of my academic family, you made Clemson my home away from home and sometimes made me forget I was away from loved ones- Dr. Jerome McClendon thank you for being a stellar lab mate, research mentor, and brother. My dissertation would have not been possible without your constant support and belief in my abilities. You also cared for my well-being, patiently listened to me complain and gave me advise that has helped me grow as a person. I aspire to be a mentor just like you. Dr. Lauren Dukes, thank you for being such an awesome friend always encouraging me to keep going. Dr. Toni Pence, I could not have ever imagined meeting someone as awesome as you. From the first day, you made me feel welcome, helped me with classes, went out of your way to check on me and indeed become one of my closest friends and someone whose advice I take unquestioningly. Dr. Ellie Ebrahimi, my confidant,friend, and real-life superwoman. You walked this journey with my hand in hand, and my mission is more meaningful for having made a friend like you. We have shared success, failures laughs and tears and it's a bond which I will treasure for life. John 'jay' Porter, you filled the void my closest friends left when they graduated. Although you are younger than me in age, your maturity, wisdom, and love have been a pillar of support for me. You are a lifelong ally I am blessed to have, thank you for being there for me unconditionally. Divine and Alex thank you for being my loudest cheerleaders your constant support and love mean the world to me. I would like to thank my grandmother,father and mother for their love and prayer that has enabled me to complete this journey. To my baby sister, thank you for stepping up to the plate and taking care of the family, so I did not have to worry about it. Sushir thank you for your support through all these years, I am truly privileged to be able to call you my best friend and partner.

Most of all, I " praise God Almighty" for having shown me the path and blessing me with the love and support of all my friends and family.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Sierra Landry from Lancaster (South Carolina) started dating Tanner Crolley and within a few months she had isolated herself from her family, dropped out of high school and her parents started noticing bruises, which kept increasing in size. Sierra tried to get away from this abusive relationship but her controlling boyfriend continued harassing and stalking her. On December 30, 2013, she went out of the house to meet her friends and a couple of hours later her body was discovered in an isolated yard. She had been shot in the head, and the local police found her boyfriend trying to flee the town and arrested him [81].

Ortralla Mosley, a high school student from Reagan High School in Austin, Texas was stabbed to death by her running back boyfriend, Marcus McTear, in March 2003. Marcus was an extremely popular athlete but he had a history of violent behavior, which led to him stabbing his girlfriend six times in the school hallway when she tried to break up with him [57].

In May 2010, a similar story unfolded again but this time in a college setting, where Yeardley Love, a sophomore lacrosse player from the University of Virginia, was found dead in her apartment. Her boyfriend George Huguely confessed to having a bitter altercation and shaking her repeatedly resulting in her head striking against the wall multiple times [15]. Brittany Henderson, a freshman at the University of Wisconsin, attended a speech on dating abuse and realized that she was in an abusive dating relationship. She was dating a popular football star but he kept analyzing her appearance and reminding her that many girls wondered why he had chosen to be with "someone like her". This often escalated to verbal abuse and threats of physical violence using social media. Brittany's father came to her rescue and helped her out of her tragic relationship [23]. The story of Sierra, Brittany, Ortralla and Yeardley are not isolated newspaper headlines - it is the story of one in three women in United States. They are victims of dating abuse and this statistic is higher than any other youth violence related crime [81]. This dissertation is organized into seven chapters. Chapter 1 describes the motivation behind this research and background information. Chapter 2 describes the research approach and specific aims. Chapter 3 illustrates the first iteration of creating the user interface of an Android app; Chapter 4 describes the pilot study to evaluate the mockup of the app's user interface. Chapter 5 describes the use of machine learning in abuse detection, Chapter 6 describes the performance evaluation of different machine learning algorithms. Chapter 7 describes the development process of the prototype android phone app. In Chapter 8 we evaluate the usability of the phone app and discuss user trust and effectiveness. Chapter 9 describes a preliminary user study to investigate different visualization paradigms that will assist users to trust the feedback regarding the possible abusive nature of their online communication, Chapter 10 discusses limitation and future work and Chapter 11 discusses the overall research contributions.

1.2 Defining Dating Violence

Dating violence is defined as "encompassing a range of violent, abusive, or threatening behaviors including physical, psychological, or emotional and sexual violence or abuse, in addition to behaviors which may be considered as controlling or dominating toward a romantic or dating partner and that cause harm, pain or injury to the victim" [68]. Dating violence is considered digital dating abuse when the perpetrator communicates through text messages, instant messaging services and social media sites such as Facebook, Twitter, or Snapchat. The categories of digital dating abuse are similar to that of dating violence.

1.3 Defining Categories of Dating Abuse

Physical violence is the most visible form of abuse in an abusive relationship. The victims face serious consequences like severe bruises, or broken bones and such injuries can result in the loss of life. Physical abuse is the red flag that friends and family notice and this is also often the final phase of abuse. Thirty-two percent (32%) women in United States have reported experiencing physical abuse from age fourteen (14) to their college years [79]. The National Longitudinal Study of Adolescent Health conducted a survey among seven thousand (7000) high school students; they found that ten percent (10%) of young women have reported being pushed around by a romantic partner and three percent (3%) reported having something thrown at them [36][38]]. Rape, attempted rape and coercive sex are all forms of sexual abuse. These terms point towards sexual relationships where the victim does not give consent or is coerced to give consent to get out of dangerous situations like beating or physical abuse from the perpetrator.[37] Financial control is another form of abuse, though more prevalent in cases of domestic violence. The abuser, in this case, may stop the victim from getting a job or completing school. This in turn leads to the victim being dependent on the abuser financially. In some cases where the victim is earning, it has been reported that their abusers have taken control over their finances [34][16]. Emotional abuse is the invisible form of abuse, which often goes undetected and this makes it an easy and lethal weapon. Hoffman, et al., defined emotional abuse as "behavior sufficiently threatening to the woman so that she believes her capacity to work, interact with family or society or to enjoy good physical or mental health, has been or might be threatened" [40] [67]. Insult, humiliation, use of abusive language, in either public or private, all constitute emotional abuse. This usually results to a lowered self-esteem, extreme confusion and severe depression. Common forms of emotional abuse include yelling, name-calling, verbal intimidation and constant check on the partner's whereabouts. Some perpetrators stalk their girlfriends, read their emails or private messages from other people and go to the extent of monitoring them even when they talk to family or close friends. According to the Center for Disease Control and Prevention (CDC), each year twenty-five percent (25%) of adolescents report experiencing "verbal, physical, emotional, or sexual abuse" from a dating partner [32].

1.4 Dating Abuse Related Statistics

Nearly 1.5 million high school students nationwide experience physical abuse from a dating partner in a single year. Girls and young women between the ages of sixteen to twenty four (16-24) experience the highest rate of dating abuse, almost triple the national average. Among female victims of dating abuse, a current or former boyfriend or girlfriend has victimized ninety-four percent (94%) of those aged sixteen to nineteen (16-19) and seventy-percent (70%) of those aged twenty to twenty-four (20-24). Nearly half (43%) of college women who are dating report experiencing violent and abusive dating behaviors. College students are not equipped to deal with dating abuse – fifty-seven percent (57%) say it is difficult to identify and fifty-eight percent (58%) say they do not know how to help someone who is experiencing it. One in three (36%) college students who are dating have given a dating partner their computer, email or social network passwords and these students are more likely to experience digital dating abuse. One in six (16%) college women have been sexually abused in a dating relationship [31].

1.5 Risk factors associated with abuse (Consequences)

The consequences of such violence can result in higher academic dropout rates, drug use, unwanted pregnancies, depression, suicide attempts and even fatalities. Younger women who have been in abusive relationships are at a higher risk of becoming pregnant or contracting sexually transmitted infections as they are unable to negotiate safe sex behaviors with their partners[71]. Being in a violent relationship often leads to depression, low self-esteem and can lead to suicidal tendencies. In a study, adolescent girls who have recently faced dating violence reported sixty percent (60%) more likelihood of attempting one or more suicides [63]. A study conducted in South Carolina in 2000 found that victims of dating abuse are more prone to drinking (32.7%) and doing drugs (48.5%) [7].

1.6 Demographics of the victims

Data collected from the CDC's Youth Risk Behavior Surveillance System (YRBSS), a national survey administered to high school students, shows that both male and female students reported being "hit, slapped, or physically hurt on purpose by their boyfriend or girlfriend during the 12 months before the survey" [28]. Dating violence is also more prevalent in communities of color due to their minority status, socio-economic status and cultural beliefs. Studies have suggested that African American females are at greater risk (35% higher) from dating violence and other forms of physical abuse compared to white females and twenty two percent (22%) higher than other races. The study conducted in rural counties of North Carolina found that adolescents from minority groups, single parent homes and homes where parents are less educated, reported facing more physical violence compared to others [33].

1.7 Digital Dating Abuse

Technology has changed the way in which young adults communicate. Internet services such as social media sites, instant messaging and email applications have made communication more accessible. Unfortunately, these internet services can also be used to harass people indiscriminately [54]. One such form of harassment is digital dating abuse (DDA), where a dating partner leaves repeated threat messages over the phone or social media. Studies have shown that undergraduate college students are prone to this form of aggression. One-third (1/3) of college students have reported facing some form of cyber-based harassment [82]. In the focus group studies conducted by Melander [54], themes were identified relating to the use of technology in abusing someone. The study concluded that online harassment is a form of "quick and easy violence" where "private becomes public" knowledge. Participants also said that in-person, aggressive, and private arguments could easily become an aggressive public domain conversation [17]. This can increase the threat and shame that the abused partner feels, and force them to reconcile to the terms set by the abuser. Digital dating abuse has become a major mental health crisis for the college-going population. Advancement in technology has made it easier to keep the partner 'captive' even when the abuser is not in the same physical location. Constant threats via text messages, surveillance using the GPS locations of the victim's digital footprint and control tactics has made dating violence a menace that needs to be detected and stopped at an early stage.

1.8 Relevance of Technology in Dating Abuse among Young Adults

A growing body of literature has acknowledged that electronic communication technology has had a significant effect on the lives of young adults. These methods include mobile phones, communication through text messages, instant messaging services and social media sites such as Facebook, Twitter, or Snapchat [70] . Nine in ten (9 in 10) college students use a laptop or smartphone on a regular basis. Half of them regularly use a tablet, while roughly one in ten (1 in 10) regularly use all or a mix of all [44]. Using these communication media, bullying and abuse has moved from in-person to the cyber world. Online bullying or cyberbullying amongst peers has been a topic of research for several years [27][4][3] but digital dating abuse is yet to be fully explored. The Centers for Disease Control and Prevention has recognized the relevance of technology within its definition of dating violence, whereby emotionally abusive and controlling behaviors may be perpetrated electronically, in addition to behaviors such as stalking [4]. Hate or abusive speech refers to speech which demeans or attacks a person or people as members of a group with shared characteristics such as race, gender, religion, sexual orientation, or disability [[30]. Digital Dating Abuse (DDA) is targeted specifically towards dating partners whereas hate/abusive speech and cyberbullying can be targeted towards anyone even if the abuser is unknown to the victim. DDA has recently been compared to cyberbullying as the "willful and repeated harm inflicted through the medium of electronic text" [65]. Cyberbullying tends to occur between individuals who do not like or want to be around each other, whereas DDA transpires between two people who are attracted to each other on some level [88].

1.9 Current Abuse Prevention and Awareness Programs

State and federal governments and Non-Governmental Organizations have initiated several dating abuse prevention and awareness programs. One such program is the Relationship Abuse Prevention Program that includes lesson plans, parent workshops, and staff development materials for a school-based curriculum [17].This program has been developed in New York City to teach students to recognize and prevent teen relationship abuse. Former President Barack Obama and the White House have also been very active in protesting against such violence, making a proclamation in 2013 designating February as National Teen Dating Violence Awareness and Prevention Month [3]. Along with awareness campaigns, technological interventions have been created: two such web-based educational platforms are Dating Matters and thatisnotcool.com [62][85]. Dating Matters: Understanding Teen Dating Violence Prevention is an initiative launched by The Centers for Disease Control and Prevention and is designed to help educators, school personnel, youth leaders, and others working to improve the health of teens. This is an online course with interactive visuals and scenario based learning modules that communicate the relevance of dating violence prevention. Thatisnotcool.com provides educational material related to dating violence and prevention too. It also has material to support teens going through abusive situations that motivates them to get help. Another example of technology-based intervention is Love is Not Abuse, an iPhone application launched by Liz Claiborne Inc. that is designed to provide parents with a simulation of situations common in digital abuse, such as threatening to remove friends from social networks or posting illicit photos. While the application does not actually access the user's Facebook account, parents get an insight into the controlling nature of a negative teen relationship [29]. The Apps Against Abuse Technology Challenge led by former Vice President Joe Biden and Secretary of Health and Human Services Kathleen Sibelius, was created to engage software developers to help find a solution to the alarmingly high incidences of sexual violence and abuse of female college students [34]. OnWatchOnCampus and Circle of 6, the winners of the 2011 Challenge, were developed to give college students a sense of safety when they are traveling alone [1][2][6]. These application interfaces can be used by a person to alert the authorities and loved ones about where they are if they feel unsafe. Distress messages and the exact GPS location of the person are relayed so help can reach them sooner. Though both these approaches are novel ideas, they do not cater specifically to digital dating abuse victims because dialing 911 might be the victim's last cry for help. By then, the situation may have gone so out of hand that even if she can be saved from physical harm, the mental scars might be left forever. None of the above-mentioned technical and non-technical approaches provide help directly to the victim at the onset of abuse. Abuse often begins as emotional abuse and goes undetected. A detection application, which analyzes the victim's digital content, is needed for preemptive action before the abuse escalates to physical harm. For our research, we use text messages exchanged between two partners as a mode of digital communication and use supervised machine-learning techniques to detect abusive text messages. We also want to understand how we can support the design of the application based on the results of the classifier.

This dissertation document outlines the process of design, creation and evaluation of a phone application that consists of a detection tool that flags abusive digital content exchanged between partners, an educational component that provides information about digital dating abuse, and a list of nearby resources for users to locate help.

Chapter 2

Research Approach

2.1 Target Population

The target population for this phone application are college going female students above the age of eighteen (18) years who are in a dating relationship. In the US, a woman is three to six times more likely to become a victim of violence than a man. Although women and girls are more exposed to risk of violence against them, interpersonal violence affects individuals of both genders as well as those of different sexual orientations [25].

2.2 Design Justifications

2.2.1 Mobile phone application and design process

Globally, the use of mobile phones and broadband services has been rising, especially amongst young adults. In 2015, sixty-five percent (65%) of American adults owned a smartphone [78]. Fifteen percent, (15%) of Americans aged eighteen to twenty nine years (18-29) were heavily dependent on a smartphone for online access and sixty two percent (62%) of this population had used their phone in the past year to look up information on a health condition [78]. Thus, we chose a smart phone platform to build our detection application to let resources be available at anytime, anywhere and at the convenience of the user [10]. As using a smartphone in public is perceived as "normal", people can access this tool without drawing unwanted attention. Mobile phone ap-

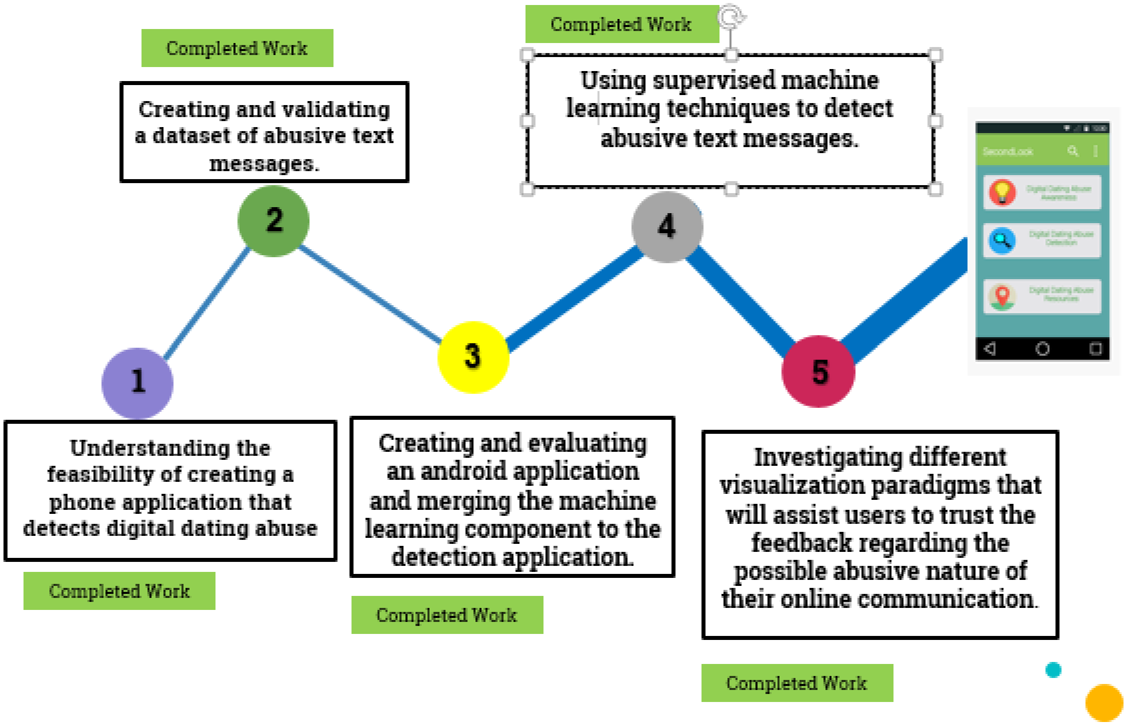

1
Understanding the feasibility of creating a phone application that detects digital dating abuse
Completed Work
2
Creating and validating a dataset of abusive text messages.
Completed Work
3
Creating and evaluating an android application and merging the machine learning component to the detection application.
Completed Work
4
Using supervised machine learning techniques to detect abusive text messages.
Completed Work
5
Investigating different visualization paradigms that will assist users to trust the feedback regarding the possible abusive nature of their online communication.
Completed Work
SecondLook
Digital Dating Abuse Awareness
Digital Dating Abuse Detection
Digital Dating Abuse Resources

Figure 2.1: Illustrates the different research phases of this dissertation, indicating completed work

plications have been used as health trackers, mental health intervention tools, and for simulating text-based dating abuse. PTSDCoach, a smartphone app for post-traumatic stress symptoms was developed by the VA's National Center for PTSD-Dissemination and Training Division. It was designed to be used as a psycho-education and self-management tool to supplement treatment by a healthcare professional [45]. Another similar example is MoodHacker, which provides effective interventions for people struggling with depression. It is a self-guided intervention tool that helps adults with mild to moderate depression learn more about strategies based on cognitive behavioral therapy to manage their mood [10]. Figure 2.1 illustrates the different phases of this dissertation.

2.2.2 Research Aims and Outcomes

My specific research aims that have been achieved are as follows:

1. Design a user-interface for the phone app using a participatory design process.

2. There is currently no publicly available dataset specific to digital dating abuse that can be

used to create training and testing sets for machine learning algorithms. Therefore, our aim is to create and validate a dataset of abusive text messages.

3. Evaluate different machine learning classification algorithms and feature extraction techniques to determine their accuracy in identifying abusive versus non-abusive text messages.

    (a) Creating an Android phone application that incorporates the results from the participatory design process to create the user interface and the detection application which uses supervised machine learning algorithms to classify text messages as abusive or non-abusive.

4. Presented with a set of text messages which contains both abusive and non-abusive text messages

    (a) Are the individual features of the app i.e. Awareness, Detection, and Resources designed in an intuitive manner that helps the user acquire the information they require?

    (b) Presented with a set of text messages which contains both abusive and non-abusive text messages

    (c) What is the threshold of abusive text messages (in percentage or count) to be considered in an abusive relationship?

    (d) Does the way the data is visualized have an impact on user's trust when it comes to using the detection application?

    (e) Is there a change in the user's perception of abusive vs. non-abusive in observing the text messages classified by the machine learning algorithm (as displayed on the app) vs. labeling the text messages themselves?

    (f) What is the threshold of abusive text messages that encourages users to seek help and if they do, which is their most preferred resources?

Chapter 3

PARTICIPATORY DESIGN PROCESS FOR INTERFACE DESIGN[1]

3.1 Introduction to Participatory Design Process

We adopted a participatory design approach for the development of our mobile application. A participatory design process is an "approach towards computer systems design in which people destined to use the systems play a critical role in designing it" [77]. Specifically, the goal of the process should be the enrichment of the user's interaction behavior; the work should be collaborative in nature and finally, the process should be iterative [14][12]. The users are at the center of the design process and are involved in every iteration of the design [56]. Thus, the users are considered as domain experts who guide the developers with technical skills to build a certain technology [24]. These ideas are very different from other conceptions of technology development where the developers are the domain experts and develop a system through a trial and error process [89]. When this process has been successfully used, it has resulted in higher user satisfaction and the application has a higher likelihood of being used [60]. This design process has also helped in creating applications with higher user productivity and better user experience [24]. While developing an application, it is essential to

[1] This work has been published in [72]

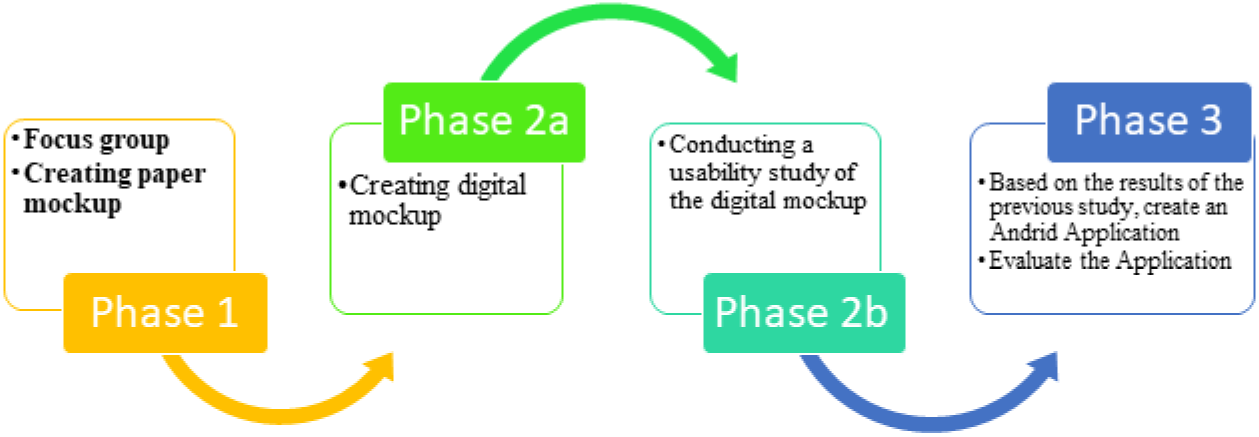

•Focus group
•Creating paper mockup
Phase 1
Phase 2a
•Creating digital mockup
•Conducting a usability study of the digital mockup
Phase 2b
Phase 3
•Based on the results of the previous study, create an Andrid Application
•Evaluate the Application

Figure 3.1: Illustrates the three phases of the participatory design process.

Age	Mean = 19.8 yrs. Standard Deviation =2.089
Gender	Female = 100%
Ethnicity	Caucasian = 90.0% African American = 10%

Education	Freshman = 60 %
	Junior =10 %
	Senior = 20%
	Master's Degree = 10%

Relationship Status	Single =100%

Table 3.1: Demographic Questionnaire Results

understand the function of the application and the needs of the user group—both of which have a high effect on system acceptance. If users are involved from the beginning of the technology design, then they also have the opportunity to develop ownership of the technology and expertise on how to use it [18]. The design goals for our application were 1) user-friendly design 2) high user-acceptance and 3) just-in-time intervention. The design process will be carried out in three phases (Illustrated in 3.1); Section 3.2. onwards describes the first phase.

3.2 Participants

For the first focus group, we recruited ten (10) female college students who were all above the age of eighteen (18) years. Recruiting women who have been in an abusive dating relationship would be ideal for getting the users' mental model. Due to the sensitive nature of this topic and limited resources, we recruited female college students as our domain experts instead of abuse victims. We conducted two focus group sessions, which were held on separate days over a duration of two hours. The researcher present with each group conducted these semi-structured sessions using open-ended questions to lead the discussion. All the discussion points were approved by the Institutional Review Board to ensure human subject research protocols were followed. For efficient data collection, the sessions were audio recorded and later transcribed into text.

3.3 Procedure

The participants were first given a demographic questionnaire to obtain their age, gender, educational status, marital status and ethnicity (Table 3.1.). Along with this, they were given a dating abuse awareness questionnaire [87], which asked about their knowledge about dating abuse. This questionnaire also asked the participants to choose from a given list, words which described signs of an unhealthy dating relationship. The remaining part of the questionnaire was about personal experiences with abuse. They were asked if they were in an abusive relationship, did they know of anyone in an abusive relationship, were they aware of resources on campus that could help them with dating abuse, did they have the friends or family they would be willing to share their ordeals with and, finally, did they want to use technology to detect digital dating abuse? After filling out the questionnaires, the researcher led a discussion with the participants. The focus group was guided by open-ended questions to encourage discussions to establish a definition of dating abuse, understand opinions on different kinds of abuse, and discuss personal experience with abuse as well as signs that can be used to recognize abuse. In the next phase part of the discussion, topics about abuse on a college campus were covered and the change in the medium of dating abuse from in-person to digital was introduced. In reference to digital dating abuse, discussions regarding medium of abuse and using technology to detect abuse were conducted. The participants also talked about their concerns regarding the use of technology to detect dating abuse. Then they were asked to draw their design ideas for the user interface for the hypothetical application. While the participants sketched their ideas, the researcher asked them questions about design choices and inspiration. The participants were not shown any user-interface designs created by the research team. This was done intentionally to reduce introducing bias into the process.

3.4 Analysis

After the completion of both the focus groups, we transcribed the audio recordings and aggregated the data into separate paragraphs depending on the focus group discussion questions. The focus group questions not only led to discussions, but lent a structure to the whole process to allow the researchers to collect quality data for further analysis. After the initial data clustering, we used the thematic analysis technique to identify themes or recurring patterns present in the data collected for the design of the user interface [13]. For this particular qualitative data analysis, the inductive or bottom up way was used to generate themes [21]. In this approach the themes are directly related to the data generated from the focus group. Inductive thematic analysis requires minimal preconception bias in the part of the researchers and the coding process does not try to label data into pre-defined categories. Themes evolve as the researchers familiarize themselves with the data and multiple iterations of data clustering. Our initial analysis process was conducted using paper and sticky notes. Later this was transferred to a digital format for further analysis. The drawings that were collected from the users were used along with the themes to create the digital mockup of the application. Quantitative data analysis was done using the demographic questionnaire and the modified dating abuse awareness questionnaire.

3.5 Results

3.5.1 Qualitative Data Analysis

[73] Several themes emerged during the focus group. Discussion about dating abuse and the use of technology in detecting it. These categories are listed below and are illustrated with sample quotes from study participants.

1. *Definition of dating abuse and its prevalence on a college campus*

All the participants agreed on the definition of dating abuse as a combination of emotional and physical abuse. Direct quotes from two participants focus on emotional or verbal abuse as being a larger threat than physical abuse.

"Sticks and stones can break your bones but words can hurt you more. Breaking the bone is huge but words can be powerful"

".. Breaking your bones can get better but words you always remember the words"

2. *Behavior changes in abuse victims*

   Participants shared stories about friends who were in abusive relationships, and how their behavior changed over a period of time. They emphasized the fact that the changes were not apparent to the abuse victim but was conspicuous to a third party. This direct quote below is a classic example of controlling behavior [5]. It also describes how a form of technology was being used for dating abuse.

   "I have never seen her without him and when in class her phone is attached to her telling him where she is."

3. *Inability to identify signs of emotional abuse*

   Though in the dating abuse awareness survey the participants answered that they knew how to identify signs of emotional abuse, during the focus group discussion this was not the case. They mentioned the dichotomy between knowing what abuse is and being unable to identify signs when it happens to them. A direct quote from one of the participants mentions that a third entity is required to have an unbiased view about their relationship.

   "Recognizing [abuse in] other people is easier, if you are the one in love it's super hard to take a step back and say ok this is actually what I want, but in other people you can see how changes happen easily"

4. *Use of technology to detect digital dating abuse* All the participants agreed that dating abuse through digital media such as text messages and social media sites are extremely relevant in a college setting. During the focus group session, most participants were interested in using technology as a tool to flag abusive content.

   "They are like started getting suspicious, but they don't want to raise a red flag, might be good to have an app that you can do it anonymously and know if you want to talk to your friends."

5. *Privacy concerns*

   All the participants had privacy concerns and feared that if the abuser found this app on the partner's phone, it may escalate the situation very rapidly. They also feared that if a non-abusive partner chanced upon this app it may cause a rift between them. Privacy concerns led them to come up with ideas on how to hide the application in plain view. The benefits of mobile app interventions mentioned in Section 2.3.1, was reiterated by the participants. The direct quote below talks about just in time anonymous interventions.

   "You do not have to talk about [abuse] during any [therapy] session, and just handing it ossut can be anonymous and they can do it on their own time."

6. *Method of delivery*

   A few participants felt more secure with their data if it was integrated into the University's existing IT framework like Blackboard. Others expressed reluctance at the notion of a third party browsing through their personal communications.

   "I don't think it would be too bad, if Clemson University would have it, I trust everything like Blackboard and stuff"

3.5.2 Qualitative data analysis: Mock-ups

We collected 10 separate user mental models of the hypothetical application. As shown in Figures 3.2,3.3,3.4 the participants drew out the user interface and listed the features they would like to see in the application. To increase ownership we also asked the users to name their app, and chose a particular layout style or color scheme if they wanted to customize the app. While the users drew out the images, the researchers asked them to describe each feature and provide design justifications. This process was influenced by the "think aloud" protocol used in usability studies [88]. All the participants included the detection component in their images, but along with that, they wanted a "help" or "resources button" (Fig.3.2 and Fig.3.4) to provide hotline numbers. Most of the participant's designs were inspired by everyday technology interfaces they use.

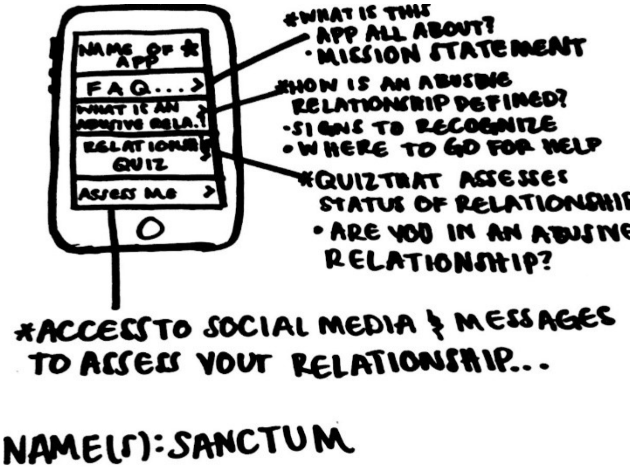

NAME OF APP
FAQ...>
WHAT IS AN ABUSIVE RELA.?
RELATIONSHIP QUIZ
ASSESS ME >
*WHAT IS THIS APP ALL ABOUT?
•MISSION STATEMENT
*HOW IS AN ABUSING RELATIONSHIP DEFINED?
•SIGNS TO RECOGNIZE
•WHERE TO GO FOR HELP
*QUIZ THAT ASSESSES STATUS OF RELATIONSHIP
•ARE YOU IN AN ABUSIVE RELATIONSHIP?
*ACCESS TO SOCIAL MEDIA & MESSAGES TO ASSESS YOUT RELATIONSHIP...
NAME(S): SANCTUM

Figure 3.2: Participant B's mock up describing a familiar user-interface style.

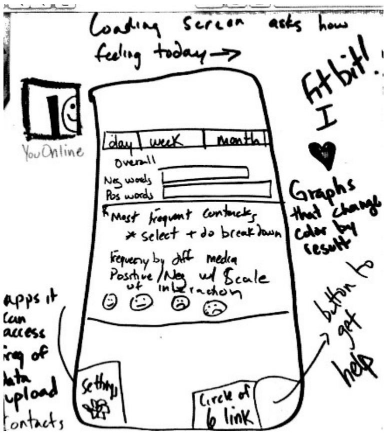


Figure 3.3: Shows a user's design closely resembling the health tracker Fitbit's user portal.

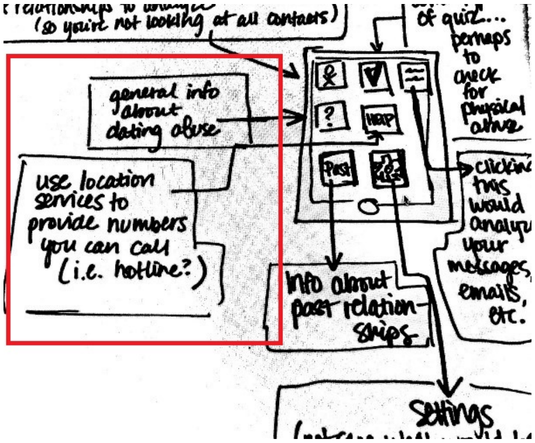

(so you're not looking at all contacts)
of quiz... perhaps to check for physical abuse
general info about dating abuse
HELP
?
Past
use location services to provide numbers you can call (i.e. hotline?)
clicking this would analyze your messages, emails, etc.
Info about past relationships
Settings

Figure 3.4: Shows Participant C's user-interface displaying three features - Awareness, resources and quiz

3.6 Quantitative Data Analysis

Figure 3.5 shows the results of the dating abuse and technology acceptance questionnaire. This questionnaire was administered before the focus group sessions.

3.7 Discussion

A different response was noticed between the results of Questions 3 and 4 from the survey and the focus group interview. In the survey, seven out of ten participants mentioned they knew what to do if they found themselves in an abusive situation, and all ten participants said they would go to a friend or counselor to confide about relationship trouble. In contrast, during the interview sessions they mentioned the difficulties in self-identifying abuse and the need of a third party intervention (Theme 3, Inability to identify signs of emotional abuse, elaborates this idea). In terms of using technology as a detection tool, eight out of ten people answered "yes" and this was validated during the focus group sessions (Theme 5, privacy concerns). In the question "Which words describe signs of an unhealthy dating relationship?" six out of ten participants chose the option "Physically violent (hitting or kicking)" and others picked words like "(coerced, possessive and name-calling)". This result was more comprehensive than the definition that the participants came up with during the focus group. All these findings further supports the development of a mobile application that can act as a complete tool for dating abuse awareness, detection and prevention. The next phase of the participatory design process is described in the next Chapter 4 and the final

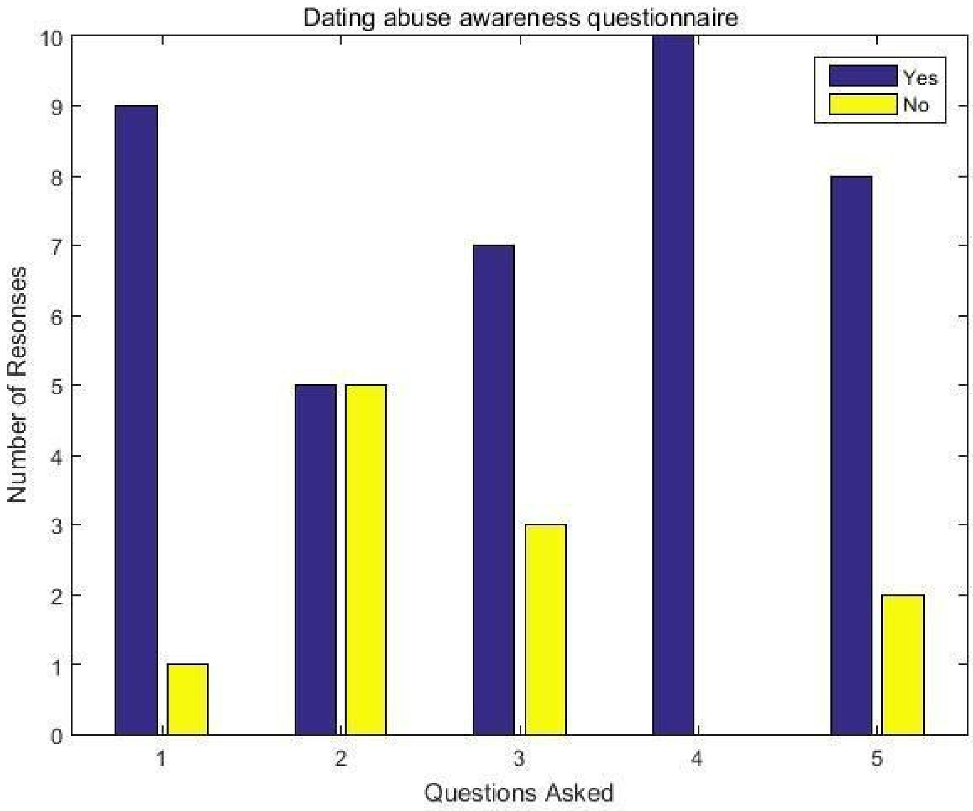


Figure 3.5: Dating Abuse and Technology acceptance questionnaire. Question 1) Have you or someone you know ever been in an unhealthy dating relationship? Question 2) Are you aware of unhealthy dating relationships at your school/ community or work place? Question 3) Do you know what to do if you find yourself in an unhealthy relationship? Question 4) Is there a friend or adult or counselor you know you can go and confide in if you were facing trouble in your relationship? Question 5) Do you want to use technology (for example - phone or web application) to help you understand if you are being abused?

phase of the participatory design process is described in Chapter 7.

Chapter 4

Pilot Study to Evaluate The Interface Design

4.1 Study design and participants

During the qualitative and quantitative analysis of the data gathered during the focus group study, we designed a low fidelity prototype of the phone application. The low fidelity prototype was developed using Fluid UI, which is a browser-based wire framing and prototyping tool and is used to design mobile touch interfaces. The purpose of the low fidelity prototype for our development process was to develop a quick mock-up, test the interactions, and design elements with experts who have experience conducting user experience research. Five (5) participants were recruited from a graduate-level computer science class at Clemson University. We did not apply for IRB approval to conduct this study, as we wanted to use this pilot to refine the design aspects before we conducted the usability study. The participants were given tablets, which had the app simulator on it. Fluid UI's simulator allows the user to interact with the prototype as if it was a working application. This re-creation of real-life interaction was essential for us to get accurate results related to usability.

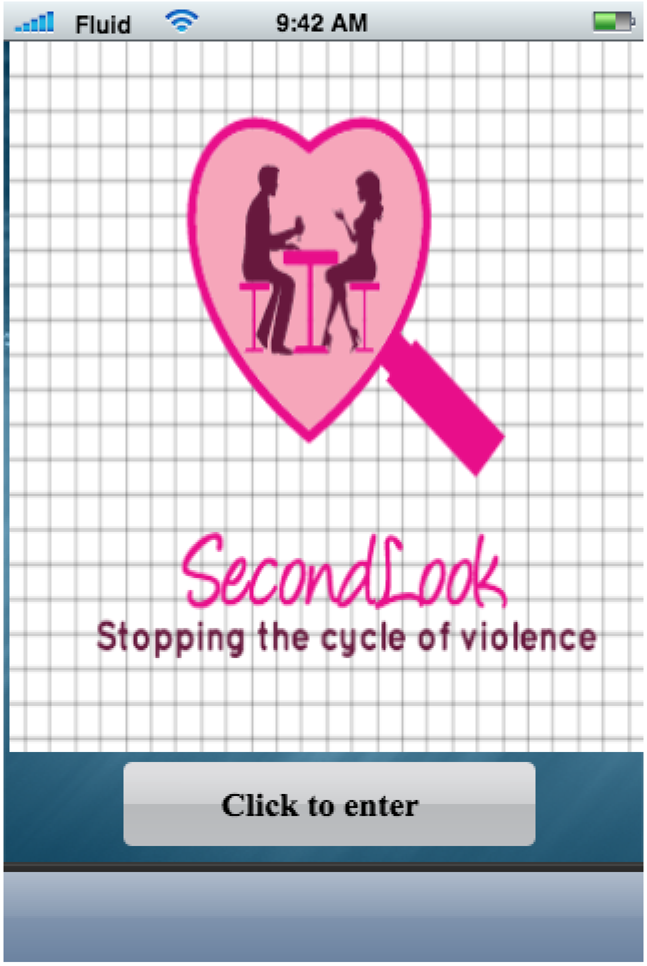


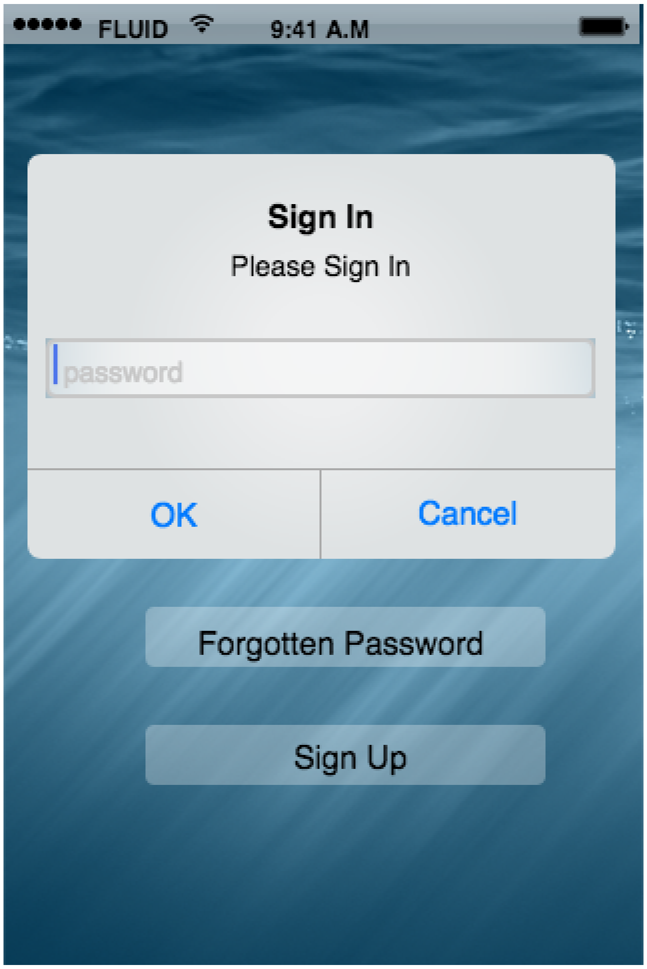
Figure 4.1: Login screen

4.2 Heuristic evaluations

4.2.1 Scale used

We performed our usability testing on this mock-up illustrated in the Figures below (Fig 4.1- 4.8). For evaluation, we used the Heuristic evaluation technique [59]. The goal of heuristic evaluation is to find usability problems in an existing design. Our participants where are asked to do a series of tests on the user interface and determine the flaws in it. For our usability testing, we used selected, our participants who had intermediate levels of doing usability testing.

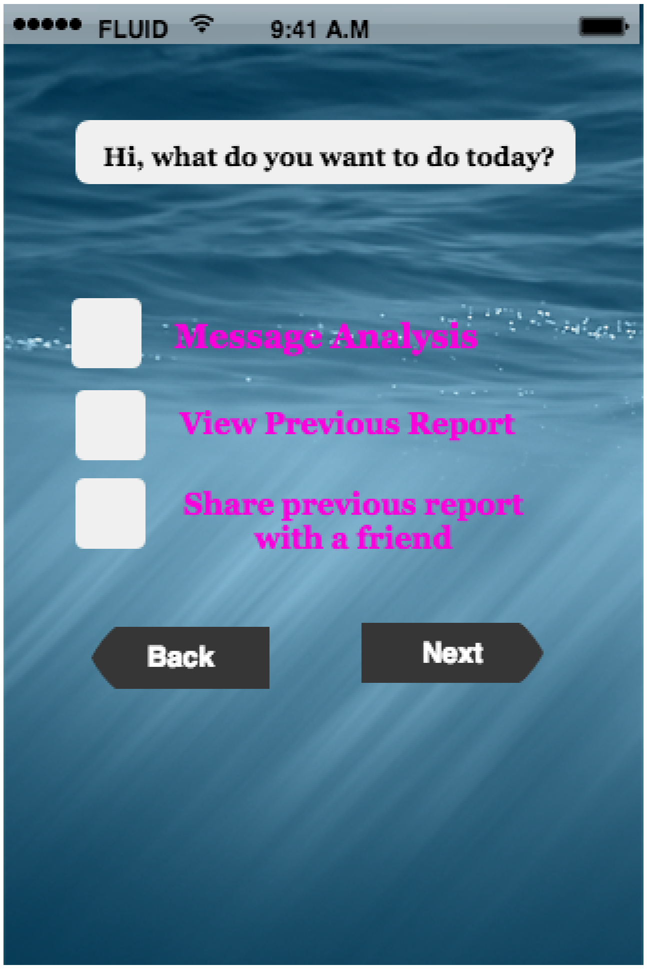
The participants were asked to carry out three different tasks using the FluidUI simulator. After the completion of the tasks, the users were asked to rate the severity of the issues using a five



Sign In

Please Sign In

Figure 4.2: Sign in / Sign up Screen



Hi, what do you want to do today?

Figure 4.3: Homes Screen

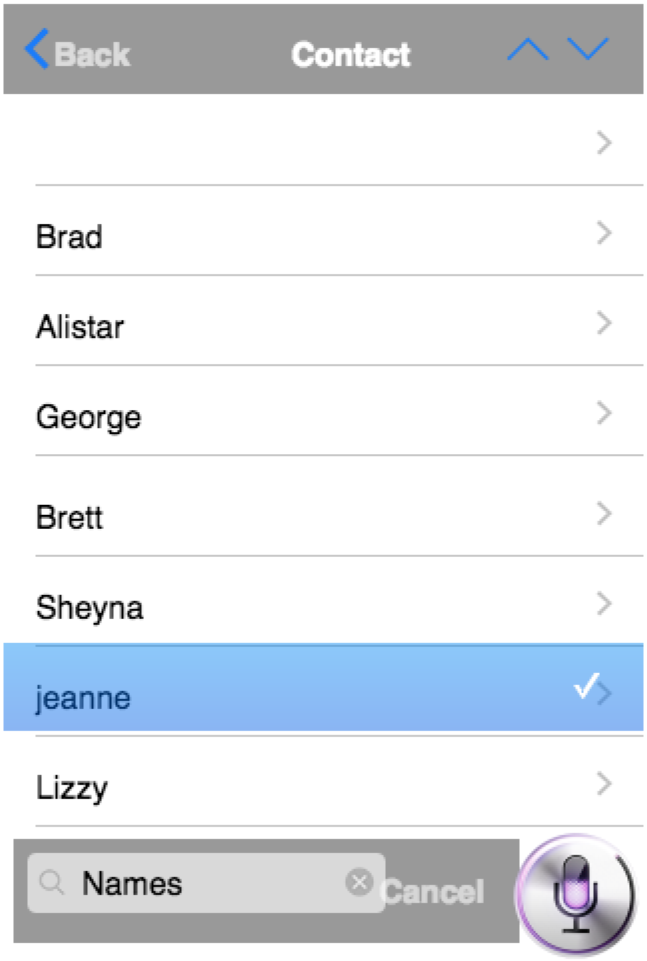


Figure 4.4: Choosing a contact

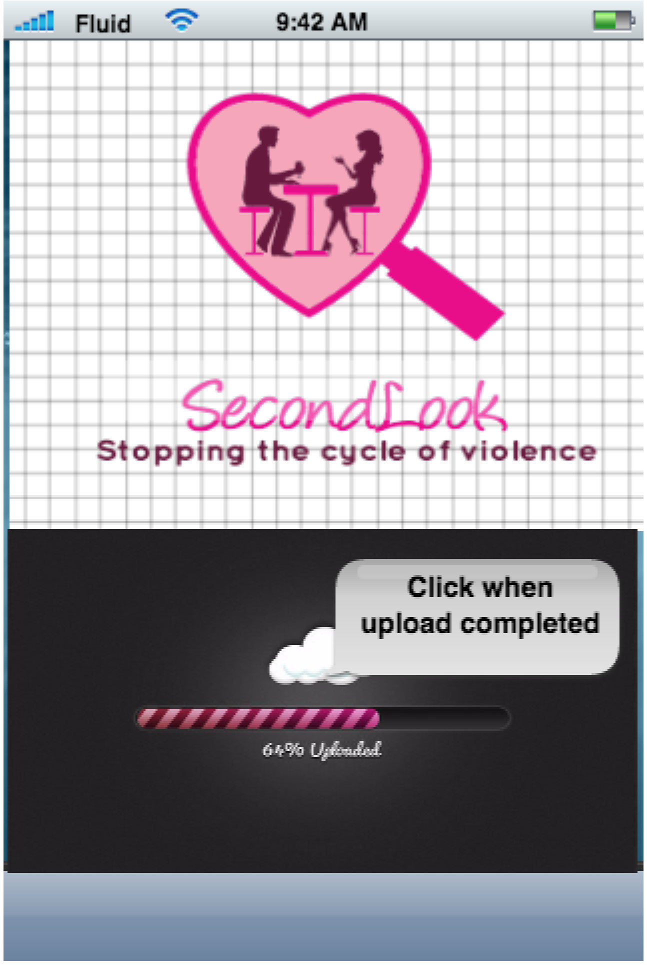
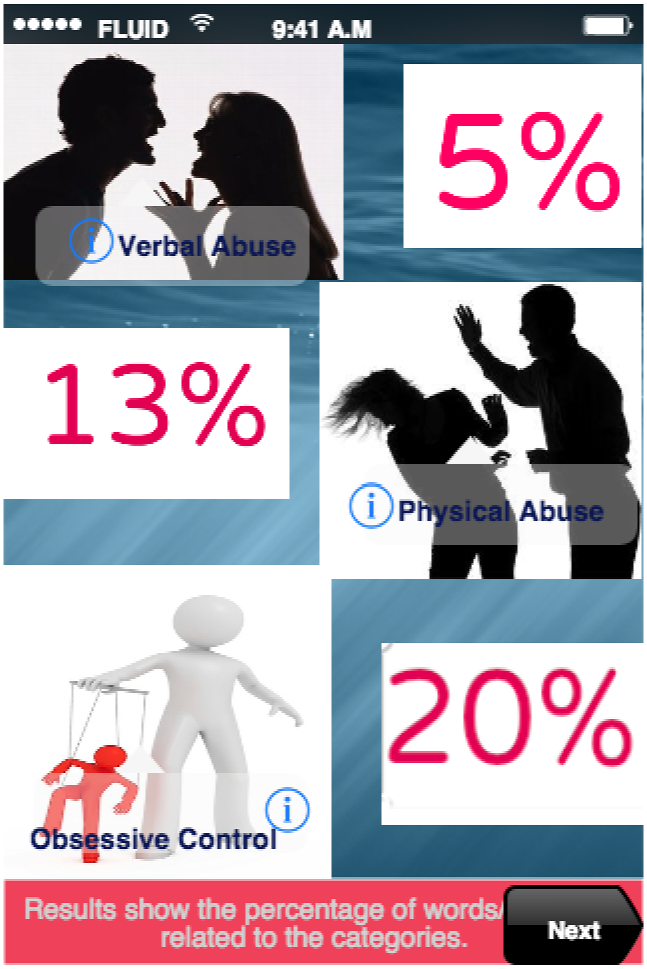
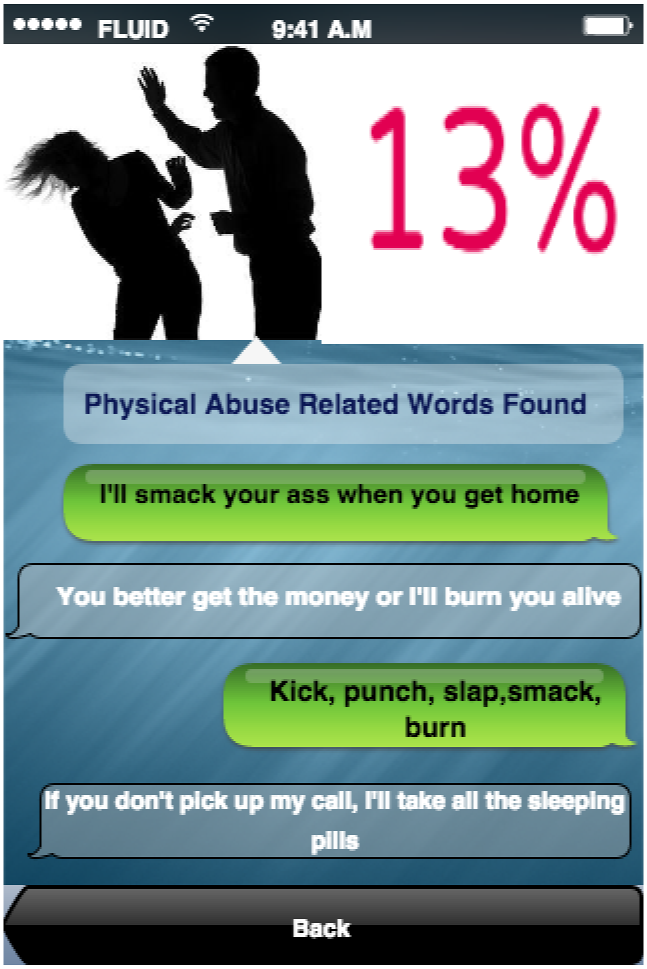


Figure 4.5: Uploading all text messages



Results show the percentage of words related to the categories.

Figure 4.6: Displaying the results of the analysis



Physical Abuse Related Words Found

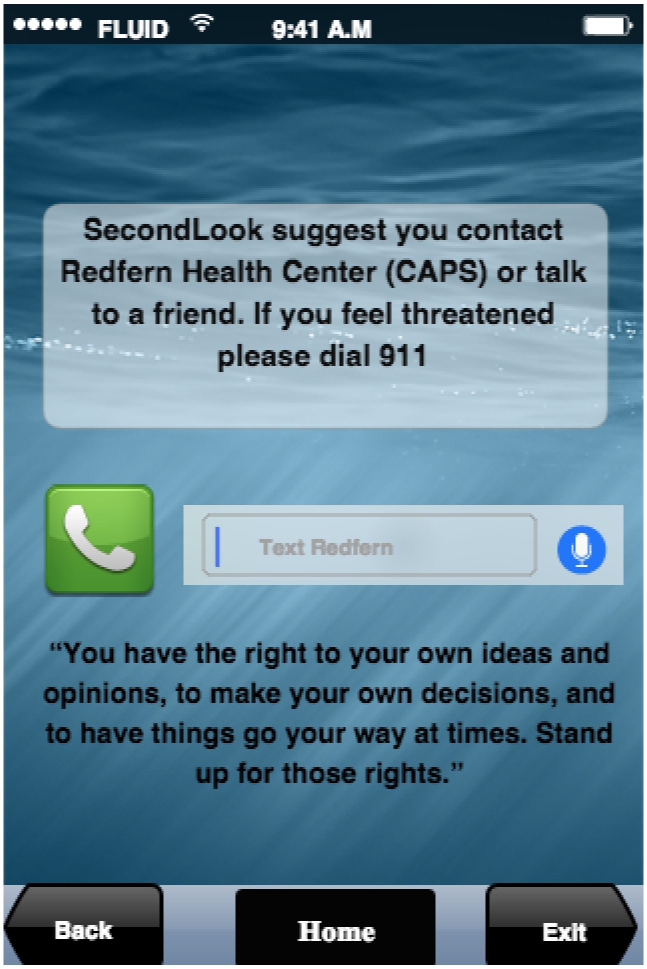
I'll smack your ass when you get home

You better get the money or I'll burn you alive

Kick, punch, slap,smack, burn

If you don't pick up my call, I'll take all the sleeping pills

Figure 4.7: Shows the abusive phrases



SecondLook suggest you contact Redfern Health Center (CAPS) or talk to a friend. If you feel threatened please dial 911

“You have the right to your own ideas and opinions, to make your own decisions, and to have things go your way at times. Stand up for those rights.”

Figure 4.8: Displays the nearest resource users could avail

scale severity rating.

4.2.2 The heuristic categories tested were [86][58][55]

1. Visibility of System Status: The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.

2. Match Between system and the real world: The system should speak the users' language, with words, phrases, and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.

3. User Control and Freedom: Users often choose system functions by mistake and will need a marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.

4. Consistency and standards: Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.

5. making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for the use of the system should be visible or easily retrievable whenever appropriate.

6. Flexibility and efficiency of use: Accelerators – unseen by the novice user – may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.

7. Aesthetic and minimalist design: Dialogues should not contain information that is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility

   The scale used was a five-point severity scale where -

   0 = I do not agree that this is a usability problem at all

   1 = Cosmetic problem only: need not be fixed unless extra time is available on the project

   2 = Minor usability problem: fixing this should be given low priority

   3 = Major usability problem: important to fix, so should be given high priority

   4 = Usability catastrophe: imperative to fix this before the product can be released

	Evaluation Categories						
Expert users	**A**	**B**	**C**	**D**	**E**	**F**	**G**
Expert 1	1	3	1	0	1	0	0
Expert 2	3	1	0	0	0	0	2
Expert 3	3	0	0	0	0	2	2
Expert 4	3	0	2	1	0	0	2
Expert 5	3	1	1	1	1	0	3

Table 4.1: Heuristic evaluation raw score

4.2.3 Tasks given

1. Log into the system and choose a contact whose conversation you want to analyze.

2. Log into the system, choose to analyze messages and view the results and read the physically abusive words aloud.

3. Go back to the home screen. Next, send the results to a friend. After completing, that return to the home screen and view previous reports. After viewing, the reports exit out of the system.

4.2.4 Results from the Evaluation

Most of the evaluators rated the system to have one major usability problem, and this was related to visibility of system status and minimalist design. The users were not being informed about where they were in the system through appropriate feedback. Specific terms used in the screens were ambiguous, such as the "Home Screen." The user was confused between the initial welcome screen (Figure 4.1) and the menu option (Figure 4.3). During task 3 when the users were asked to share the generated reports with a friend, they could not complete the task successfully as Fig. 4.2 and Fig. 4.5 both have share with a friend option, which led to confusion. Fig.4.6, all the users had difficulty understanding and following the instructions of clicking the image to find the words and phrases related to individual forms of abuse. The pink colored labels were misleading to most users as they thought it corresponded with the picture below. None of the users understood that the images displayed were a clickable button and not a static image. Users all suggested that navigating from Fig.4.6 to Fig. 4.3 which is the main control page is time-consuming as they have to use the back button several times to reach there. Users also disliked the checkerboard pattern of text and image present in Fig. 4.6 and reported that it was very distracting. Other minor usability concerns raised were related to the way the buttons were being displayed and the text-heavy labels on them. Users reported concerns over the lack of a sign-up page for new users. When the user was using the system for the first time, they should not have the option of sharing or viewing reports with a friend as there are no previous reports. Overall, the users found the system interactive and beneficial for detecting dating abuse.

4.3 User Interface Changes

We made two design changes to the user interface after our heuristic evaluation. The two primary usability flaws were visibility of the system and minimalist design. We incorporated the home and exit button options on all pages to allow the user to exit the application whenever they wanted. The "home screen" was labeled as so, to avoid the confusion that the users expressed. Finally, the wording of the labels was rewritten to give clear instructions without redundancy. This user-interface design only tested the detection feature of the phone app, after analyzing the data from the focus group study described in Chapter 3 we changed our design direction to create a phone app with features which promotes dating abuse awareness and provides the user with resources that they can use for help.

Chapter 5

Using Machine Learning to Detect Abuse [1]

5.1 Use of Machine Learning in Abuse Detection

Supervised and unsupervised machine learning algorithms have been used to identify patterns in texts. These techniques have wide-ranging uses from SPAM detection to identifying internet sexual predators. McGhee et al. used machine learning to classify online posts from online forums as sexual predation. They developed phrase matching and rule-based systems to identify appropriate features for their supervised learning algorithms. Using a decision tree and an instance-based learning algorithm they achieved a sixty-eight percent (68%) accuracy [53]. Dinakar's work on text-based cyber bullying looked at classifying comments from YouTube videos involving sensitive topics related to race, culture, sexuality, and intelligence. They performed two text classification experiments, first training binary classifiers to label a topic as sensitive or not and then a multiclass classifier that classifies a comment into a specific label. JRip (Repeated Incremental Pruning to Produce Error reduction), Naïve Bayesian, J48 (decision tree based classifier based on the C4.5 method proposed by Quinlan) and Support vector machine algorithms were used as classifiers. Their results showed that binary classifiers trained on individual topics performed better than multiclass classifiers. The rule-based JRip classifier gave the best accuracy of approximately eighty (80.20%) percent [27]. De

[1] This work has been published in [74]

Chowdhury et al., in their work in detecting depression from social media posts used Twitter as a tool for measuring and predicting depression in individuals. They focused on creating a depression lexicon from the Mental Health forum on Yahoo! Answers. A support vector machine was used to predict the likelihood of depression and the classifier had a seventy (70%) percent accuracy rate [20]. 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. A perceptron algorithm, support vector machine and neural network classifiers were used. A ninety percent (90%) accuracy rate was achieved with Linear Support Vector Machine [76] [75]. Our research in the domain of digital dating abuse is closely related to the work mentioned above. The researchers have successfully used machine-learning algorithms in detecting issues related to mental health and interpersonal violence. As digital dating abuse falls under the domain of interpersonal violence, we predicted that we could use similar machine learning methods to conduct our studies.

5.2 Text Messages and Sociolinguistic Patterns

The type of language that people use when they communicate with each other using computer-mediated communication is usually very different from speech-based communication [84]. Including abbreviations, initializations and short forms are common characteristics of text-based conversation. For example, laughter can be expressed in several ways, such as lol, lmao or ha-ha. [84]. To use Natural Language Processing (NLP) tools, it is essential to understand the semantics used and incorporate them into the training set.

5.3 Approach

Fig. 5.1. illustrates the research approach we took to address these aims.

5.4 Dataset Creation

Due to the lack of a publicly available dataset, we first had to create a dataset of abusive text messages. In this section, we describe the methodology used to create that dataset.

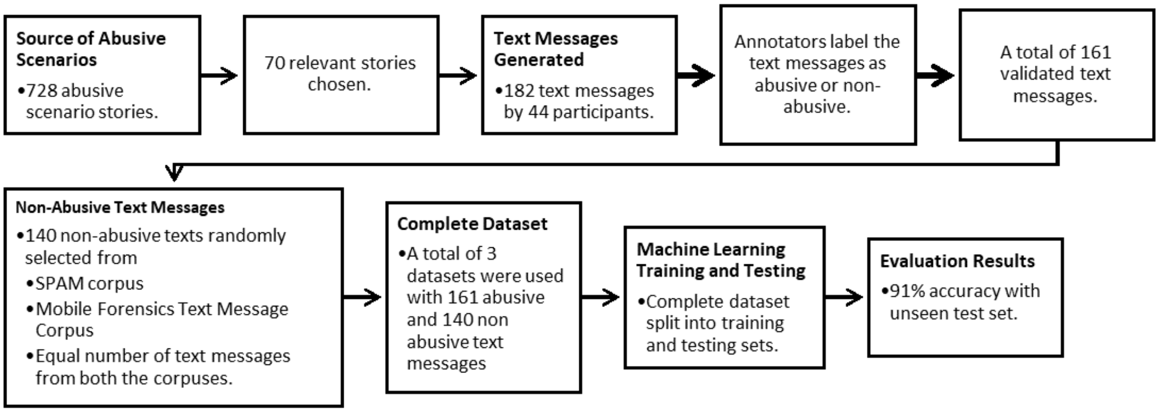

Source of Abusive Scenarios
•728 abusive scenario stories.
70 relevant stories chosen.
Text Messages Generated
•182 text messages by 44 participants.
Annotators label the text messages as abusive or non-abusive.
A total of 161 validated text messages.
Non-Abusive Text Messages
•140 non-abusive texts randomly selected from
•SPAM corpus
•Mobile Forensics Text Message Corpus
•Equal number of text messages from both the corpuses.
Complete Dataset
•A total of 3 datasets were used with 161 abusive and 140 non abusive text messages
Machine Learning Training and Testing
•Complete dataset split into training and testing sets.
Evaluation Results
•91% accuracy with unseen test set.

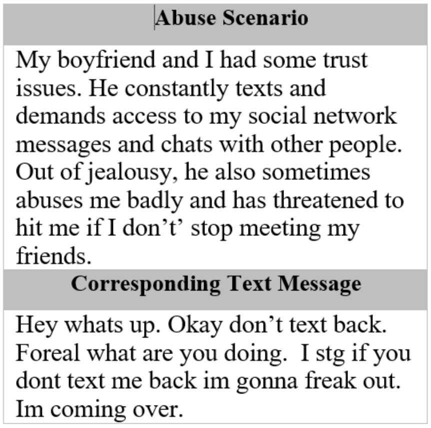
Figure 5.1: Flowchart illustrating research approach

5.4.1 Source of Abuse Scenarios

MTV's website, athinline.org is a crowdsourced feedback and advice forum for distressed young adults [62]. Teenagers and adolescents anonymously share stories, and when a story is posted, everyone can read it and vote on the severity of the story. There are three levels - over the line (severe), on the line (moderate to mild) and under the line (not very serious). The topics including sexting, online harassment, social network bullying, dating abuse, physical abuse and sexual abuse. We analyzed an anonymous corpus of 728 of these personal stories and specifically picked 70 stories that were closely aligned to different scenarios of dating abuse. These stories were then used as a starting point to create scenarios to motivate the creation of abusive text messages.

5.4.2 Participants and Procedures

We recruited forty-four (44) participants from Clemson University. Each participant was given a demographic questionnaire, followed by a dating abuse awareness questionnaire to establish a knowledge level baseline. The demographic questionnaire asked about their age, gender, ethnicity, educational background and relationship status. The demographic distribution is shown below in Table 5.1. Participants were then allotted five stories each from the abuse scenarios and asked to create abusive text messages by pretending to be the abusive romantic partner. This activity was conducted using Google forms, and the participants typed out their responses. They were also given instructions to create responses as if they were texting someone so that we could capture unique



Abuse Scenario

My boyfriend and I had some trust issues. He constantly texts and demands access to my social network messages and chats with other people. Out of jealousy, he also sometimes abuses me badly and has threatened to hit me if I don't' stop meeting my friends.

Corresponding Text Message

Hey whats up. Okay don't text back. Foreal what are you doing. I stg if you dont text me back im gonna freak out. Im coming over.

Figure 5.2: Abuse scenario story and resulting text message

linguistic patterns and short forms that young adults use while texting. A total of 170 responses were collected. Fig. 5.2. is an example of an abuse scenario story and the resulting text message created by a participant. Some of these responses contained multiple text messages, which were split into individual units of 182 text messages. Before the participants created their messages, they were verbally instructed (during the informed consent stage) on what DDA is and were shown examples of abusive text messages. They were also given instructions about including general characteristics such as acronyms, spelling and grammatical errors. The researchers then reviewed the messages for their relevance before giving them to the domain experts for validation. Although some of our participants reported being single, we believe the language structure and word usage would be similar to that of abusers in the same age group.

5.5 Dataset Validation

5.5.1 Annotators

The dataset of text messages created in Section 5.4. was given to a group of four annotators. The criteria for choosing these annotators were that they either were victims of abusive relationships, family members of victims of interpersonal violence, or they were researchers in this domain.

Age	Mean = 20.5 yrs. Standard Deviation =2.089

Gender	Male = 87.36 %
	Female = 12.3%

Ethnicity	Caucasian = 70.45%
	African American = 20.45%
	Asian/ Pacific Islander = 6.48%
	Hispanic = 2.62%

Education	Bachelor's Degree = 15.91%
	Freshman= 15.91%
	High school graduate = 15.91%
	Junior =2.27%
	Master's Degree = 2.27%
	Senior = 11.36%
	Some college credit = 18.18%
	Sophomore = 18.18%

Relationship Status	Single = 23.26% In a relationship= 76.74%

Table 5.1: Illustrates the demographic data for all the participants

5.5.2 Procedure

The annotators were each given 182 text messages. The annotators then labeled each text message as either abusive or non-abusive. If a message was judged abusive, they were then asked to rate their confidence concerning the type of abuse or abuses it represented on a scale of 1 to 5. A one indicating that the abusive text entirely did not belong to that category and five indicating that the abusive text unquestionably belonged to the category. The annotators were also provided with standard definitions of each abuse categories. The definitions of the categories of dating abuse provided to the annotators were as follows [4]:

- Physical abuse - Occurs when a partner threatens to pinch, hit, shove, slap, punch, or kick a partner.

- Physiological /Emotional abuse - Threatening a partner or harming his or her sense of self-worth. Examples include name-calling, shaming, bullying, embarrassing on purpose, or keeping him/her away from friends and family.

- Sexual abuse - Forcing a partner to engage in a sex act when he or she does not or cannot consent. Can be physical or non-physical, like threatening to spread rumors if the partner refuses to have sex.

- Stalking - Refers to a pattern of harassing or threatening tactics that are unwanted and cause fear in the victim.

5.5.3 Calculating Inter-rater Agreement

After the annotators labeled the text messages as abusive or non-abusive and classified abusive text messages into different categories of abuse, we calculated joint agreement between the annotators using Light's Kappa, an inter-rater agreement statistical consistency measure (5.1).

$$K = \frac{\overline{P_0} - \overline{P_e}}{1 - \overline{P_e}} \tag{5.1}$$

Light's kappa [49] is designed for studies with three or more annotators. Light suggests computing agreement for all coder pairs then using the arithmetic mean of these values to provide an overall index of agreement. The factor $1 - \overline{P_e}$ gives the degree of agreement that is attainable above chance, and $\overline{P_0} - \overline{P_e}$ gives the degree of agreement actually achieved above chance. The statistic takes values between 0 and 1, where a value of 1 means complete agreement. For this dissertation, we only report the agreement between annotators for the task of labeling sentences as abusive or non-abusive. The overall agreement between annotators when annotating the text messages for our study was 0.58, which in this context is considered as a moderate overall agreement [35][46]. For our final dataset of text messages, we chose 157 (of 182) text messages that were labeled as abusive by at least three out of four annotators. In four cases where there was a tie between annotators as to whether or not the text was abusive the researchers broke the tie by labeling the text as abusive. The final dataset contains 161 abusive text messages. As this app addresses a highly sensitive issue such as inter-personal violence the accuracy of the classifier is essential to help the user trust this app and use it to prevent escalation of violence. Thus training the machine learning algorithm with a diverse set of abusive and non-abusive training data was essential. The next chapter (Chapter 6) discusses the three different datasets we use for training and testing purposes and evaluation metrics and results which prompted us to choose one classifier over others.

Chapter 6

Classifiers and Performance Evaluation

[1] To train a machine-learning classifier to recognize abusive versus non-abusive text messages we created a balanced dataset of abusive and non-abusive text messages. Two different corpora were used to collect non-abusive text messages: the SMS Spam Corpus v.0.1 [22] and the Mobile Forensics Text Message Corpus [61]. To validate consistent classifier performance we created three different datasets for training and testing purposes. The text message combinations were as follows:

- SPAM - One hundred and sixty-one (161) abusive messages with one hundred and forty (140) non-abusive text messages from the SPAM corpus
- Critical - One hundred and sixty-one (161) abusive messages with one hundred and forty (140) non-abusive text messages from the Mobile Forensics Text Message Corpus
- Mixed - One hundred and sixty-one (161) abusive messages with seventy (70) non-abusive text messages from the Mobile Forensics Text Message Corpus and seventy (70) non-abusive text messages from the SPAM corpus.

We then divided each of these datasets (301 text messages) into a training set (70%) and a testing set (30%). To derive features that could be used by a classifier we used the python-based Natural

[1] This section has been published in [74]

Language Toolkit to clean, tokenize and lemmatize the text messages and then used two separate feature extraction methods [66].

6.1 Cleaning Text

We removed extra spaces, carriage returns and HTML tags from the text messages.

6.2 Tokenization and Lemmatization

Sentences can be decomposed into words, multi-word expression, contractions (e.g., couldn't), non-alphabetic characters such as symbols and emoticons. An automated process of breaking up strings into word-like units is called tokenization. Another common step when using natural language data is to lemmatize all tokens. The process of lemmatization converts tokens to their base dictionary form. In doing so, dimensionality reduction is achieved, which can help to improve application performance [9]. We used both tokenization and lemmatization to break up our training sentences into individual units.

6.3 Feature Extraction

We used two different feature extraction methods, countvectorizer and tf-idf [69]. Countvectorizer converts a collection of text documents to a matrix of token counts. Tf-idf is a common term weighting scheme in information retrieval, often used in document classification, that is computed as term-frequency times inverse document-frequency. Using tf-idf instead of the raw frequencies of occurrence of a token in a given document scales down the impact of tokens that occur very frequently in a given dataset and that are hence empirically less informative than features that occur in a small fraction of the training corpus.

6.4 Classifiers and n-gram input

The three different classifiers used were: Linear Support Vector Classification with a linear kernel and penalty parameter C= 0.1 of the error term Multinomial Naïve Bayesian Classifier Decision Tree All three classifiers were implemented using Python's sci-kit learn libraries [29]. We then tested each classifier with different input variations of unigrams, bigrams, and trigrams. In natural language processing (NLP) an n-gram is composed of n words that appear beside each other. For example, if the input sentence is “where are you now”. Unigrams would be ‘where’, ‘are’, ‘you’, and ‘now’. Bigrams would be ‘where are’, ‘are you’, and ‘you now’. Trigrams would be ‘where you are’ and ‘are you now’ [52].

6.5 Evaluation

To evaluate the performance of the classifiers we used six different performance metrics: Accuracy, Confusion Matrix, Receiver Operating Characteristic Curve(ROC-AUC score and ROC – curveplot) , Precision, Recall and F1 score. All these performance metrics were implemented python’s sklearn metrics package [66].

6.6 Accuracy

We calculated the accuracy for each of the three classifiers and compared it to combinations of unigram, bigram and trigram inputs and two separate feature extraction methods. If y'_i is the predicted value of the i-th sample and y_i is the corresponding true value, then the fraction of correct predictions over n-samples is defined as the accuracy score (6.1). Accuracy scores range between 0 and 1.

$$accuracy(y, y') = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(y'_i = yi) \tag{6.1}$$

5-fold cross-validation [43] statistical analysis was conducted on the training set, and average accuracy for every fold was calculated. After the classifiers were trained, we used an unseen test set to evaluate performance. Accuracy was calculated for every unseen test set across all the three datasets (Critical, SPAM and Mixed). Figure 6. 1 (5-fold cross validation) and Figure 6.2 (unseen testset) illustrates the accuracy results for each of the classifiers across all the datasets .

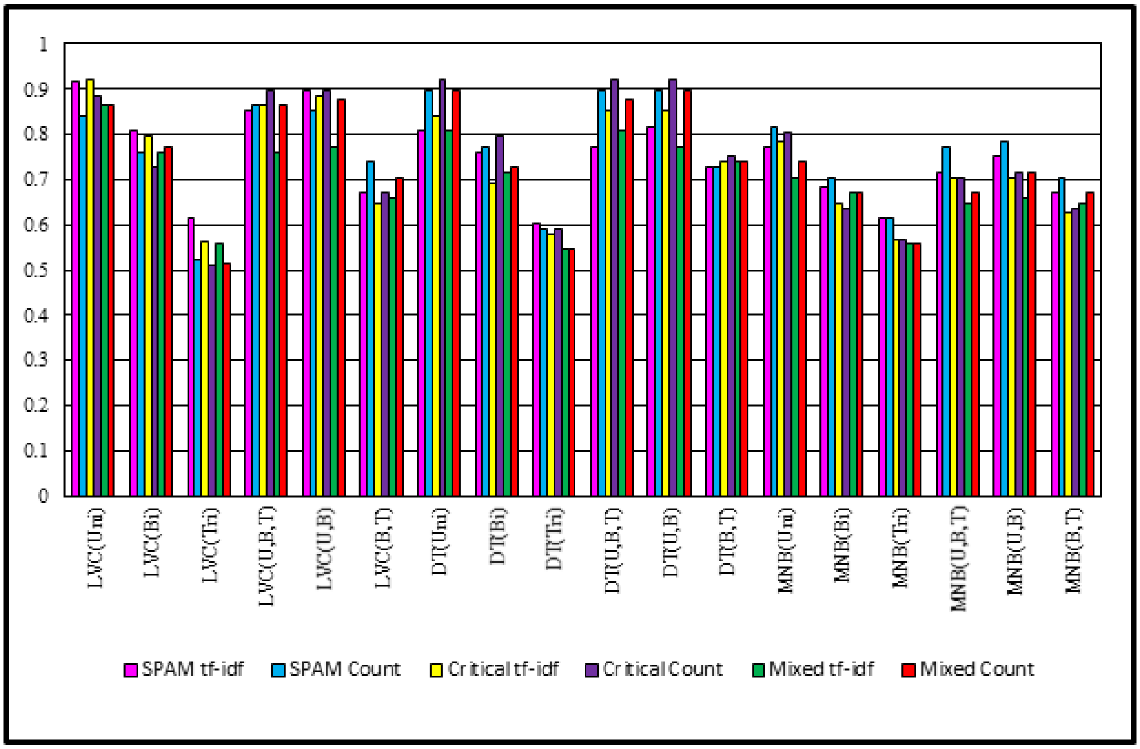


Figure 6.1: Illustrates average classifier accuracy for 5-fold cross validation of all the different classifiers, n-gram input and feature extractors across three different datasets

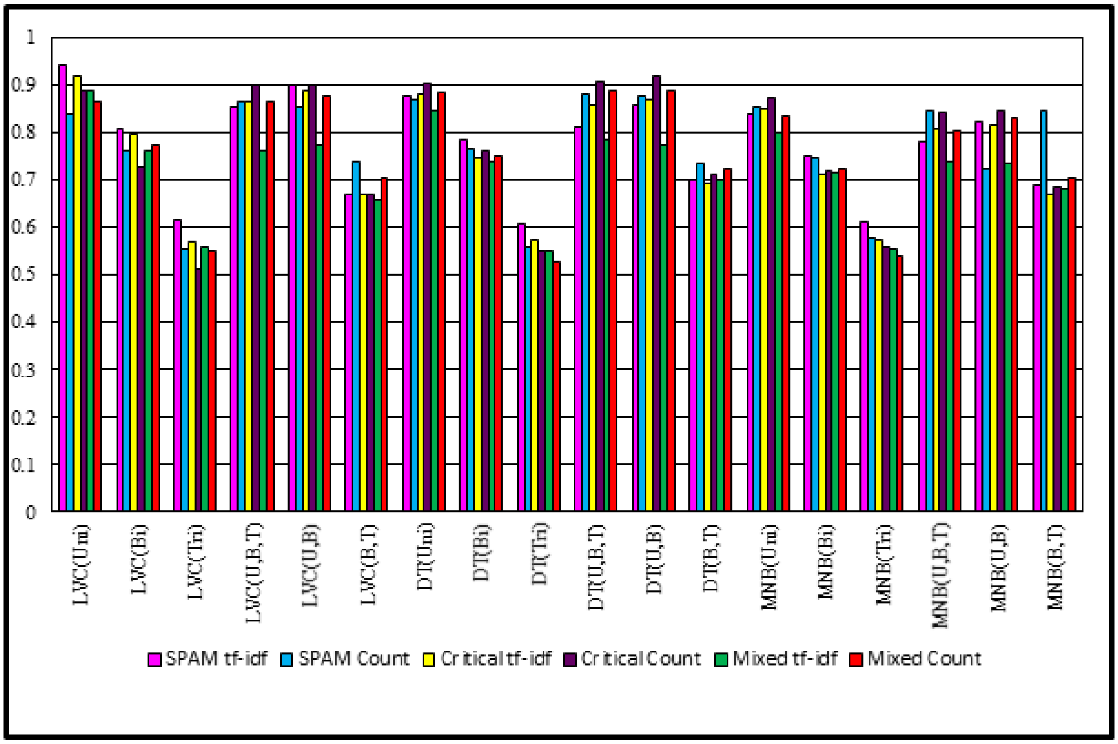
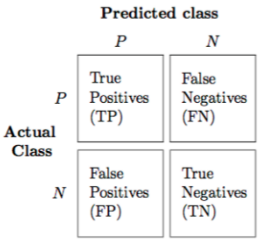


Figure 6.2: Illustrates classifier accuracy for unseen test set for all the different classifiers, n-gram input and feature extractors across three different datasets



	P	*N*
P	True Positives (TP)	False Negatives (FN)
N	False Positives (FP)	True Negatives (TN)

Figure 6.3: Illustrates the structure of a confusion matrix

6.7 Confusion matrix

We also used a confusion matrix to evaluate the accuracy of classification. A confusion matrix is a 2 by 2 array that reports the number of false positives, false negatives, true positives and true negatives produced by the classifier. This allows for a more detailed analysis than just the number of correct classifications. Fig. 6.3. Illustrates a confusion matrix. For our study, abusive text messages are considered positive, and non-abusive text messages are considered negative.

- A true positive (tp) was when an abusive sentence is classified as abusive.
- A true negative (tn) was when a non-abusive sentence is classified as non-abusive.
- A false positive (fp) was when a non-abusive sentence is classified as abusive.
- A false negative (fn) was when an abusive sentence is classified as non-abusive.

In a binary decision problem, a classifier labels examples as either positive or negative. The decision made by the classifier can be represented in a structure known as a confusion matrix or contingency table. Figure 6.4 shows the results of the confusion matrix in the format [TP, FN] [FP, TN].

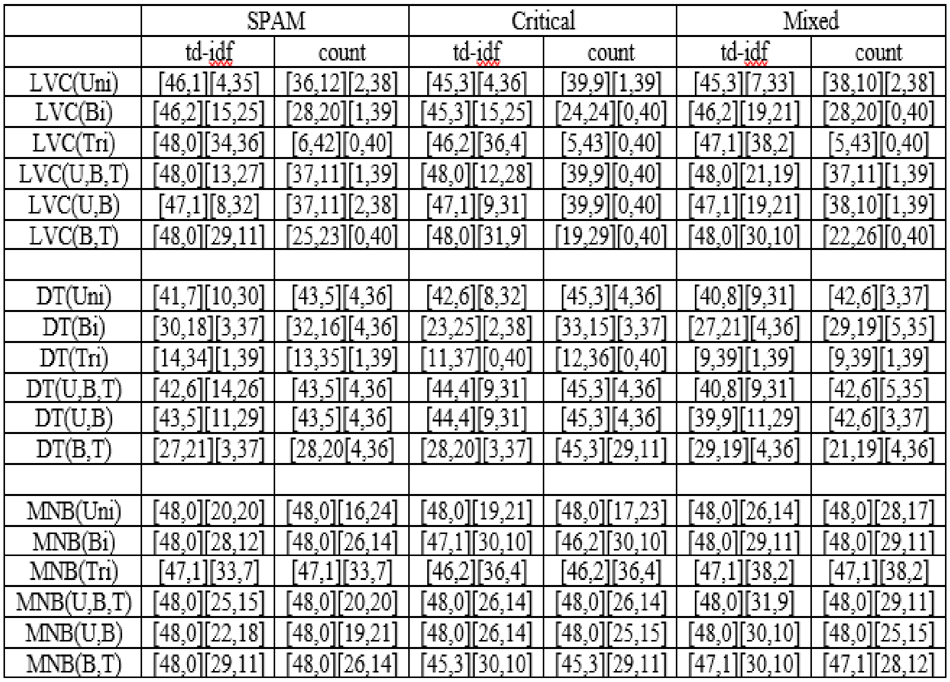
6.7.1 Precision and Recall

A false positive rate (FPR) is defined as the fraction of negative examples that are misclassified as positive. The true positive rate (TPR) measures the fraction of positive examples are labeled correctly. Recall is equivalent to true positive rate and precision measures the fraction of examples classified as positive that are truly positive [80].

$$Precision = \frac{t_p}{t_p + f_p} \tag{6.2}$$

$$Recall = \frac{t_p}{t_p + f_n} \tag{6.3}$$

For both Equations 6.2 and 6.3 as shown above *tp is true positive, fp is false positive, and tn is true negative.* Figure 6.6 and 6.7 illustrates the values of precision and recall.



	SPAM		Critical		Mixed	
	td-idf	count	td-idf	count	td-idf	count
LVC(Uni)	[46,1][4,35]	[36,12][2,38]	[45,3][4,36]	[39,9][1,39]	[45,3][7,33]	[38,10][2,38]
LVC(Bi)	[46,2][15,25]	[28,20][1,39]	[45,3][15,25]	[24,24][0,40]	[46,2][19,21]	[28,20][0,40]
LVC(Tri)	[48,0][34,36]	[6,42][0,40]	[46,2][36,4]	[5,43][0,40]	[47,1][38,2]	[5,43][0,40]
LVC(U,B,T)	[48,0][13,27]	[37,11][1,39]	[48,0][12,28]	[39,9][0,40]	[48,0][21,19]	[37,11][1,39]
LVC(U,B)	[47,1][8,32]	[37,11][2,38]	[47,1][9,31]	[39,9][0,40]	[47,1][19,21]	[38,10][1,39]
LVC(B,T)	[48,0][29,11]	[25,23][0,40]	[48,0][31,9]	[19,29][0,40]	[48,0][30,10]	[22,26][0,40]
DT(Uni)	[41,7][10,30]	[43,5][4,36]	[42,6][8,32]	[45,3][4,36]	[40,8][9,31]	[42,6][3,37]
DT(Bi)	[30,18][3,37]	[32,16][4,36]	[23,25][2,38]	[33,15][3,37]	[27,21][4,36]	[29,19][5,35]
DT(Tri)	[14,34][1,39]	[13,35][1,39]	[11,37][0,40]	[12,36][0,40]	[9,39][1,39]	[9,39][1,39]
DT(U,B,T)	[42,6][14,26]	[43,5][4,36]	[44,4][9,31]	[45,3][4,36]	[40,8][9,31]	[42,6][5,35]
DT(U,B)	[43,5][11,29]	[43,5][4,36]	[44,4][9,31]	[45,3][4,36]	[39,9][11,29]	[42,6][3,37]
DT(B,T)	[27,21][3,37]	[28,20[4,36]	[28,20][3,37]	[45,3][29,11]	[29,19][4,36]	[21,19][4,36]
MNB(Uni)	[48,0][20,20]	[48,0][16,24]	[48,0][19,21]	[48,0][17,23]	[48,0][26,14]	[48,0][28,17]
MNB(Bi)	[48,0][28,12]	[48,0][26,14]	[47,1][30,10]	[46,2][30,10]	[48,0][29,11]	[48,0][29,11]
MNB(Tri)	[47,1][33,7]	[47,1][33,7]	[46,2][36,4]	[46,2][36,4]	[47,1][38,2]	[47,1][38,2]
MNB(U,B,T)	[48,0][25,15]	[48,0][20,20]	[48,0][26,14]	[48,0][26,14]	[48,0][31,9]	[48,0][29,11]
MNB(U,B)	[48,0][22,18]	[48,0][19,21]	[48,0][26,14]	[48,0][25,15]	[48,0][30,10]	[48,0][25,15]
MNB(B,T)	[48,0][29,11]	[48,0][26,14]	[45,3][30,10]	[45,3][29,11]	[47,1][30,10]	[47,1][28,12]

Figure 6.4: Confusion matrix results of the unseen test set for all the different classifiers, n-gram input and feature extractors across three different datasets. The following notations indicate: three classifiers (LVC – Linear Support Vector Classifier, DT- Decision Tree, MNB – Multinomial Naïve Bayesian Classifier) , two feature extractor conditions (countvectorizer and tf-idf) and six input combinations (Uni – unigrams, Bi – bigrams, Tri- trigrams, U,B,T – unigrams, bigrams and trigrams, U, B – unigrams and bigrams, B,T – bigrams and trigrams). The format of the confusion matrix is [true positive, false negative] [false positive, true negative].

6.7.2 F1 Score

The third evaluation metric we used was the F1 score (also F-score or F-measure) which is another measure of a test's accuracy (Fig.6.5) [90]. F1 score considers both the precision p and the recall r of the test to compute the score: p is the number of correct positive results divided by the number of all positive results returned by the classifier, and r is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive). The F1 score is the harmonic average of the precision and recall, on a scale of 0-1, a score of 1 is considered the best F1-score The formula for the F1 score is 6.4:

$$F1Score = \frac{2}{\frac{1}{recall} + \frac{1}{precision}} \tag{6.4}$$

6.7.3 Area Under the Receiver Operating Characteristic Curve (ROC – AUC score and plot)

The final metric used was the ROC AUC Score (Fig. 6.8). A Receiver Operating Characteristic Curve (ROC) "is a plot of the true positive rate against the false positive rate for the different possible thresholds during a test. It is created by plotting the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate), at various threshold settings. TPR is also known as sensitivity, and FPR is one minus the specificity or true negative rate." On a scale of 0-1, a score of 1 is considered an excellent ROC-AUC score [39]. We also used the roc-curve function in sklearn, to plot the ROC curves and they are illustrated in Appendix A

6.8 Results

Evaluations were carried out on all three dataset combinations, to establish which classifier consistently performed better compared to the others. Fig 6.2. illustrates the results of average accuracy across all five folds in the 5-fold cross-validation tests. For all three datasets, the combination of a linear support vector machine, tf-idf, and unigrams performed better than the other classifier, feature extractor, and n-gram input combination. The accuracy values are 0.943, 0.92 and 0.8863 respectively. Amongst the unseen test sets, the best accuracy result achieved was ninety-one

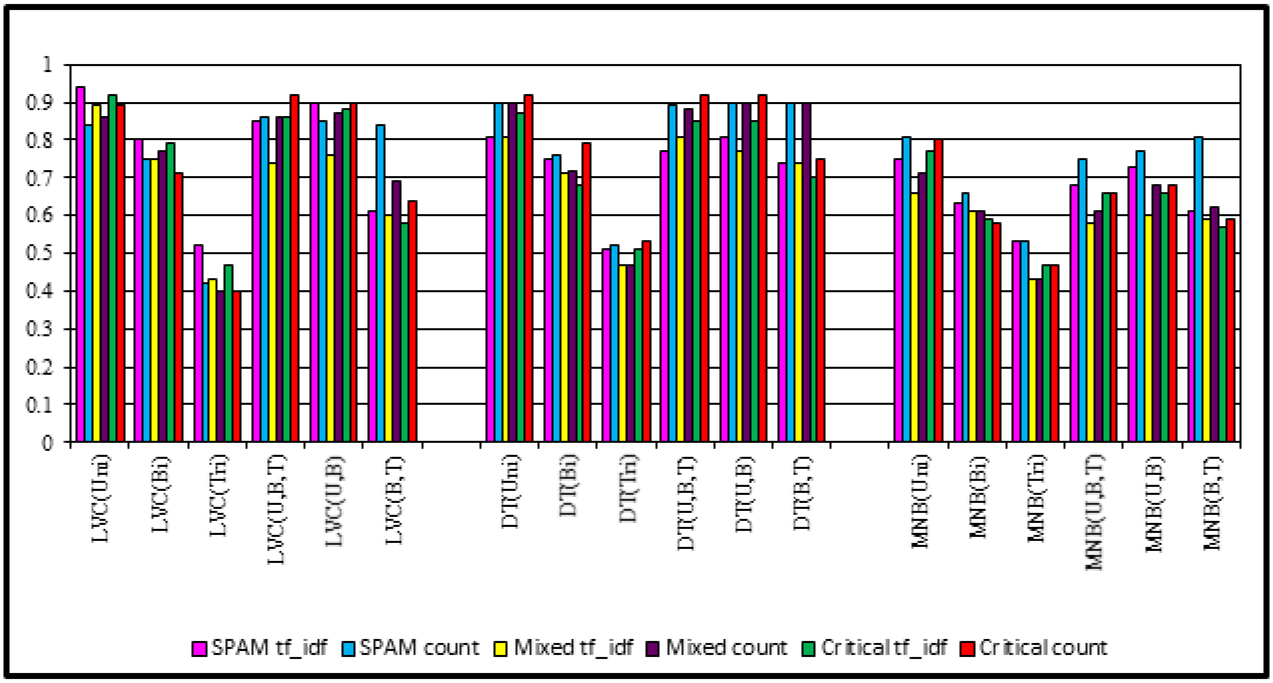


Figure 6.5: Illustrates the **F1 Score** for all the different classifiers, n-gram input and feature extractors across three different datasets.In the Figures 3a,3b and 4 the following notations indicate three classifiers (LVC – Linear Support Vector Classifier, DT- Decision Tree, MNB – Multinomial Naïve Bayesian Classifier) , two feature extractor conditions (countvectorizer and tf-idf) and six input combinations (Uni – unigrams, Bi – bigrams, Tri- trigrams, U,B,T – unigrams, bigrams and trigrams, U, B – unigrams and bigrams, B,T – bigrams and trigrams

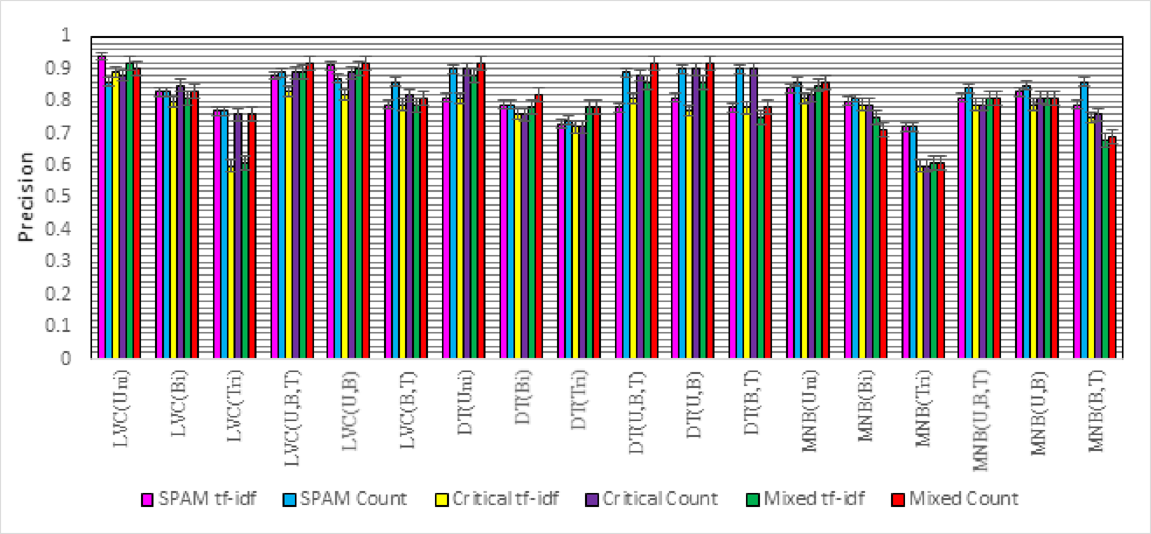


Figure 6.6: Illustrates the **Precision** for all the different classifiers, n-gram input and feature extractors across three different datasets. In the Figures 3a,3b and 4 the following notations indicate three classifiers (LVC – Linear Support Vector Classifier, DT- Decision Tree, MNB – Multinomial Naïve Bayesian Classifier) , two feature extractor conditions (countvectorizer and tf-idf) and six input combinations (Uni – unigrams, Bi – bigrams, Tri- trigrams, U,B,T – unigrams, bigrams and trigrams, U, B – unigrams and bigrams, B,T – bigrams and trigrams).

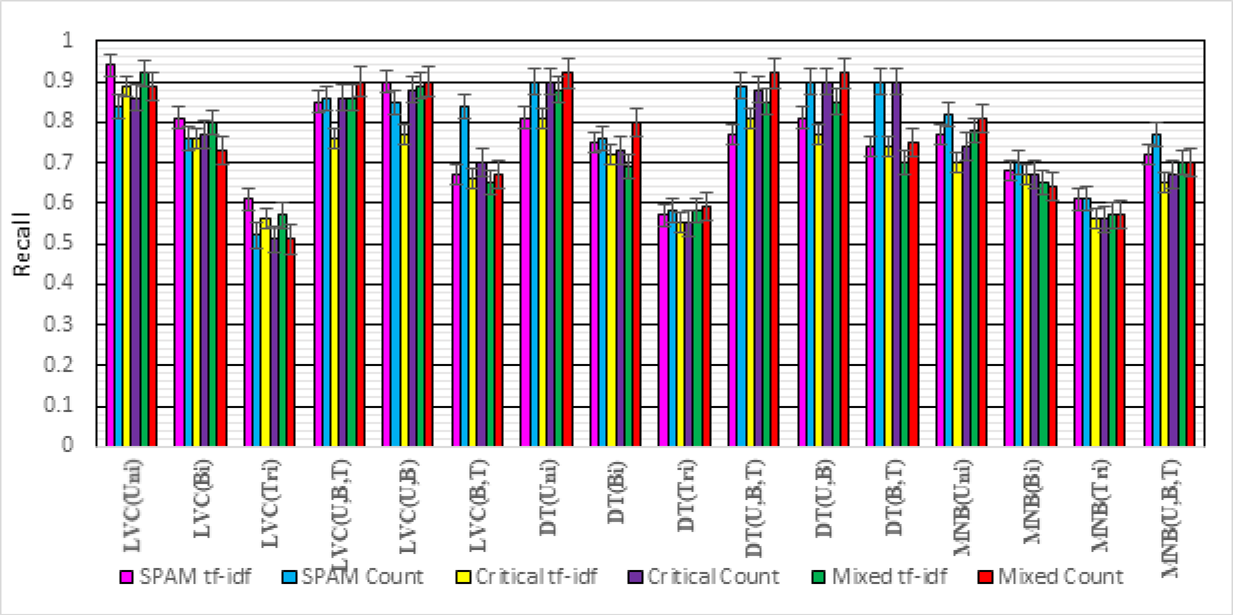


Figure 6.7: Illustrates the **Recall** for all the different classifiers, n-gram input and feature extractors across three different datasets.In the Figures 3a,3b and 4 the following notations indicate three classifiers (LVC – Linear Support Vector Classifier, DT- Decision Tree, MNB – Multinomial Naïve Bayesian Classifier) , two feature extractor conditions (countvectorizer and tf-idf) and six input combinations (Uni – unigrams, Bi – bigrams, Tri- trigrams, U,B,T – unigrams, bigrams and trigrams, U, B – unigrams and bigrams, B,T – bigrams and trigrams).

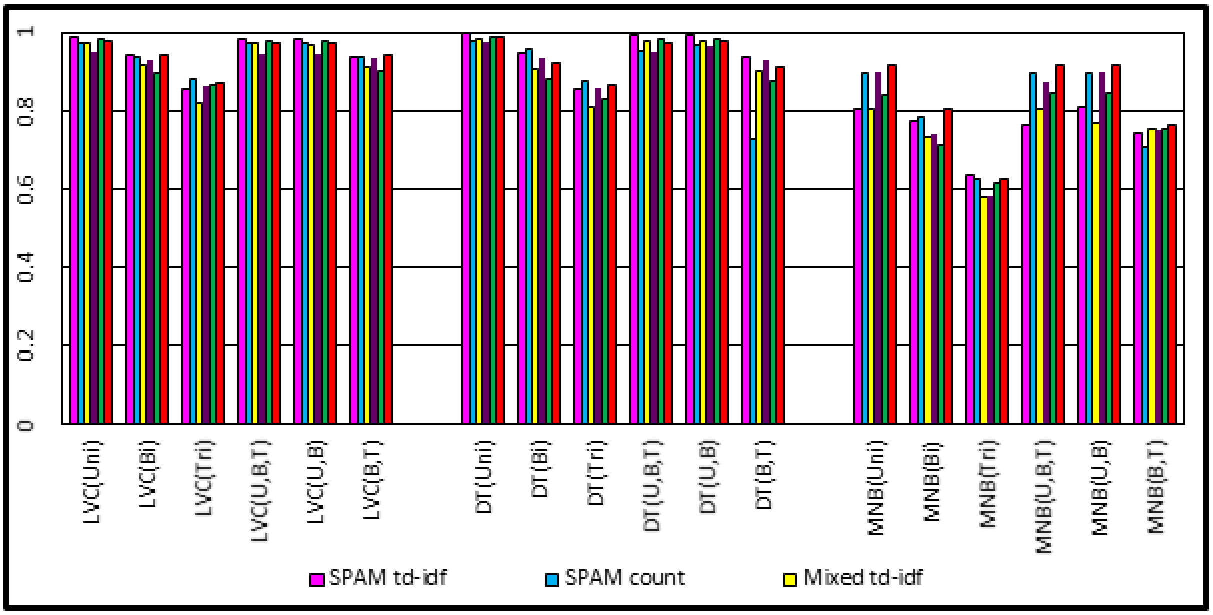


Figure 6.8: Illustrates the **ROC AUC Score** for all the different classifiers, n-gram input and feature extractors across three different datasets. In the figure the following notations indicate : three classifiers (LVC – Linear Support Vector Classifier, DT- Decision Tree, MNB – Multinomial Naïve Bayesian Classifier) , two feature extractor conditions (countvectorizer and tf-idf) and six input combinations (Uni – unigrams, Bi – bigrams, Tri- trigrams, U,B,T – unigrams, bigrams and trigrams, U, B – unigrams and bigrams, B,T – bigrams and trigrams).

percent (0.9166) using unigrams with tf-idf feature extractor and a linear support vector classifier. This result is similar to the accuracy achieved by the researchers whose work was discussed in the related works section. Fig 6.1. illustrates the accuracy values of all the different classifiers, n-gram input and feature extracts across three different datasets. An ideal configuration for a confusion matrix is when the number of true positive and true negatives conforms to the actual input. For our data we would like to see numbers approximating forty-eight (48) true positives and forty (40) true negatives, indicating that both abusive and non-abusive text messages have been labeled correctly. The best balance of correctly identifying both true positives and true negatives was obtained by using unigrams with the tf-idf feature extractor and a linear support vector classifier on the unseen test sets. This was consistent for all three variations of the dataset, and these results support the accuracy values described above. Table 6.2 illustrates the results from the confusion matrices for all the datasets across all combinations. To further validate our results, we calculated the F1-score and ROC AUC scores which have been described in Section 6.5.5 [75]. Fig 6.4. illustrates all the F1 scores with the highest value of 0.94 being achieved with a Support Vector Machine, a unigram input, and a tf-idf feature extractor when we used the SPAM dataset. In case of the other two datasets (Critical and Mixed), the values for the combination mentioned above was 0.92 and 0.89 respectively. The ROC AUC score (Fig. 6.5) was 0.9895 in case of the SPAM dataset - a Support Vector Machine, a unigram input, and a tf-idf feature extractor. The ROC AUC scores for the other two datasets (Critical and Mixed) were 0.98 and 0.979 respectively for the same combination of classifier, feature extractor and n-gram input (Support Vector Machine, unigram input, and tf-idf feature extractor). [76]. Across the three datasets the evaluation metrics do not have identical values, but for all of them, the combination of a linear support vector machine classifier, unigram input, and tf-idf feature extractor outperform other classifiers. Although we have found no other published studies in this area of research, we can compare our accuracy with studies carried out in the broader domain related to interpersonal violence prevention and mental health. Across all three datasets, the accuracy percentage has ranged from approximately eighty-eight to ninety-one percent (88-91%). This range of accuracy is closely aligned with the performance evaluation data collected in the studies that are described in the related work section. The list below shows the the binary classification results of the combination of a linear support vector machine classifier, unigram input, and tf-idf feature extractor on the unseen test set. Each text message is followed by a ':' and then a label 'a' or 'na' which indicates abusive and non-abusive respectively.

1. Why won't you message me back? Call me or the tape goes on the internet. I have the tape you whore. How do you think your family will like the tape bitch? :a
2. why won't you respond to me?! Text me back bitch know you got these. Stop avoiding me!why don't you want to talk! I can't :a
3. stand this I'm going to find you and then we'll see how much talking you do :a
4. WTH.Bitch.Asshole :a
5. Do thisfor me.DO it.WTH :a
6. You are an annoying bitch and I cannot believe that I believed what you told me. :a
7. gimme ur fb pw right now you crazy girl. right now. who r u tslking to you? tell me ASAP :a
8. Who are you always texting? I know that you're cheating? If I catch you don't send me the passwords, I'll kill you and who ever you are tweeting. :a
9. Send me nudes or i will hit you. titsz please :a
10. If you do not send me nudes then I will slander your name :a
11. Im tired of your shit Dana. If you don't get back with me then I will force you too :a
12. send me nudes.why u not answering me,if u dont fuckin send me nudes i'll tell everyone how much of a prude u are u little bitch :a
13. Im not wearing any rubber and you should understand where im coming from itll be better for both of us. If you ever tell em that i have to wear a rubber to fuck ill tell everyone of FB you have the clap. Youre completely fucking stupid if you think im wearing a condom just to fuck you. :a
14. I always knew you were a whore. Now I have proof you're a whore, and your dad will know you're a whore too....unless you think it's about time we start having sex. That dude can kiss you on the first night and I can't fuck after months? :a
15. Sometimes I thought ""you know what, she's a good girl, no need to pressure her"", but now I know that is the furthest thing from the truth so it's about time to feel some pressure. Sex tonight or I show this picture to your dad and you know how much he likes me. :a

16. Who do you think you are?! Cheating on me?! I've never met a girl with that kind of nerve. To be honest, I need to feel like I'm special, so I say you come over tonight and we have sex or else this secret kiss will be in your dad's messages tomorrow. :a

17. Come on. Just one more time. You better come over or I will spread things about you that you will not like. :a

18. hey what you doing? you talking to some other guy rn? wtf is wrong with you i love youcome on babe dont be like that wtf this isnt cool im your bf, not those guys :a

19. wtf how could u betray me like that? quit looking at other dudees y arent you replying to me? fuck u im sending those pics to your sororit ywtf is wrong with you why dont you wanna come over? i know youre not sleeping im gonna send those pics :a

20. Hey I saw you hanging out with that girl the other day. What the hell!? You would rather spend time with a girl instead of your man. you must be high cause your making stupid choices. No lesbian is gonna treat me like a piece of s****t. :a

21. Hey girl send me some nude pics. Everyone is doing it I don't see why you can't. Nothing is wrong with it it's not like i'm gonna show it to someone else. Like what the hell i touch you all the time and when i ask you to send me a freakin nude pic you start acting like a b***h. I'm comin over there right now to set you straight. :a

22. Why the f***k are you calling me. You don't need to know what I'm doing all the time, that is information only I need to know.So shut the hell up and stop texting me you b***h. I don't have time for you s***t tonight. :a

23. I know you like me. Come one just let me have a feel. It's ok you may not like it now but trust me you will. No one else is :a

24. gonna give you the same amount of attention I do so just let me do what I want. You think your some prize?!? That some other guy is gonna treat you like princess. Well get over yourself if you want to stay with me you better start following my rules. :a

25. Girl what the hell happened to you. You looked fine before but now what the hell. Your so freaking fat. Have you looked in a mirror? Oh i'm sorry i shouldn't have asked because you

obviously haven't. I bet the other guys are thankful that i saved from you. Man your lucky IM sticking this out to see where this goes. :a

26. ay shawty you paid the bill yet.I know you got money. Go ahead and pay the bill.You wasting my damn time. Wait til I get home :a

27. Where the heck are you? dont be lying to me. you dont ever tell me anywhere you aregoing. I am going to start following you around if you dont tell me. :a

28. Youre so fucking stupid the ass was OK at best but im gonna show all my homies. Youre a pretty decent booty call, but making me wait was a complete pain in the ass so im gonna show the whole school how good you are. I cant believe you really thought that after i fucked i would stay youre so stupid, im showing everyone the video. Dating you was a complete waist of time so i hope you enjoy becoming a porn star. :a

29. i dont want to see you talking to other guys".i should be the only one you want to talk to".you have no reason to be seeing people over than me :a

30. Hey whats up. Okay don't text back. for real what are you doing. I stg if you dont text me back im gonna freak out. Im coming over. :a

31. give me your facebook password now.I'm not gonna stop annoying you until you give me your password.if you don't give it to me right now i am gonna hack your facebook :a

32. What you up to babe.Babe?Where are you right now?Probably having fun with guys. Alright thats cool im going dt tonight and what happens with other girls happens :a

33. Send me nudes or i will hit you. titsz please :a

34. Cunt.Bitch.Slut :a

35. lets hook up tonight. cmon. i know you want to.you better, or else. :a

36. Bitch didnt I tell yo ass I was wit the boys. Stop texting my muthafuckin phone I will be there when I get there. Dont let me have to tell you again or imma beat yo ass. :a

37. Wtf you need to know for. Ill be there when I get there. You asking me questions aint going to get me there no faster so leave me tf alone before I break for fuckin neck. :a

38. Gtfoh I told you where I was you need to chill before I slap yo ass again. sick of you textin my phone with this bullshit. Dont let me have to repeat myself. :a

39. WTH.Slut.Whore :na

40. you dont know any better than to just fool around. you are dumb and dont get it :a

41. Why are you acting like a scared little bitch? It's not like I haven't seen it before. I'll just tell everyone I've seen you naked anyways, so you might as well send it. :a

42. why I gotta use a condom.If you loved me then you wouldnt make me wear one If I have to wear a condom I'm gonna tell everybody that youre a dub bitch that has a chlamydia and you want me to wear one because you want to keep fucking :a

43. and you dont want me to find out :a

44. I want to hold you all the time. Do you miss me? I want to be with you all the time.Text me back NOW. I NEED TO HEAR FROM YOU Please call me. I need to know how you feel. Let me look into your eyes :a

45. hey baby I saw you winking at me in the bar, why don't we get out of her so I can make you smile I want you and I'm sure you want me to, give me your address so we can want each other together I'm bet you miss my touch, why don't you come over here so I can give you what you need :a

46. Hey baby, wheres your sexy ass been? I know you want this dick girl dont try to be hard to get. I saw that ass today in your leggings;) you know i wanna fuck. Bitch stop playing and come get this dick. :a

47. What are you doing? WTH Bitch :a

48. You wanna see something *eggplant emojiSend me a pic tho. Nah a different type of pic. I already know what your face looks like.Why you taking so long to text back. You trying to be with somebody else other than ya boy :a

49. Please don't text me anymore. I have nothing else to say. :na

50. Okay name ur price as long as its legal! Wen can I pick them up? Y u ave x ams xx :na

51. I'm still looking for a car to buy. And have not gone 4the driving test yet. :na

52. As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune :na

53. wow. You're right! I didn't mean to do that. I guess once i gave up on boston men and changed my search location to nyc, something changed. Cuz on my signin page it still says boston. :a

54. Umma my life and vava umma love you lot dear :na

55. Thanks a lot for your wishes on my birthday. Thanks you for making my birthday truly memorable. :na

56. Aight, I'll hit you up when I get some cash :na

57. How would my ip address test that considering my computer isn't a minecraft server :na

58. I know! Grumpy old people. My mom was like you better not be lying. Then again I am always the one to play jokes... :a

59. Dont worry. I guess he's busy. :na

60. What is the plural of the noun research? :a

61. Going for dinner.msg you after. :na

62. I'm ok wif it cos i like 2 try new things. But i scared u dun like mah. Cos u said not too loud. :na

63. GENT! We are trying to contact you. Last weekends draw shows that you won a -1000 prize GUARANTEED. Call 09064012160. Claim Code K52. Valid 12hrs only. 150ppm :na

64. Wa, ur openin sentence very formal... Anyway, i'm fine too, juz tt i'm eatin too much n puttin on weight...Haha... So anythin special happened? :na

65. As I entered my cabin my PA said, " Happy B'day Boss !!". I felt special. She askd me 4 lunch. After lunch she invited me to her apartment. We went there. :na

66. You are a winner U have been specially selected 2 receive -1000 or a 4* holiday (flights inc) speak to a live operator 2 claim 0871277810910p/min (18+) :na

67. Goodo! Yes we must speak friday - egg-potato ratio for tortilla needed! :na

68. Hmm...my uncle just informed me that he's paying the school directly. So pls buy food. :na

69. PRIVATE! Your 2004 Account Statement for 07742676969 shows 786 unredeemed Bonus Points. To claim call 08719180248 Identifier Code: 45239 Expires :na

70. URGENT! Your Mobile No. was awarded -2000 Bonus Caller Prize on 5/9/03 This is our final try to contact U! Call from Landline 09064019788 BOX42WR29C, 150PPM :na

71. here is my new address -apples& pairs& all that malarky :na

72. Todays Voda numbers ending 7548 are selected to receive a $350 award. If you have a match please call 08712300220 quoting claim code 4041 standard rates app :na

73. I am going to sao mu todayẆill be done only at 12 : na

74. predict wat time 'll finish buying? :na

75. Good stuff, will do. :na

76. Just so that you know,yetunde hasn't sent money yet. I just sent her a text not to bother sending. So its over, you dont have to involve yourself in anything. I shouldn't have imposed anything on you in the first place so for that, i apologise. :a

77. Are you there in room. :na

78. HEY GIRL. HOW R U? HOPE U R WELL ME AN DEL R BAK! AGAIN LONG TIME NO C! GIVE ME A CALL SUM TIME FROM LUCYxx :na

79. K..k:)how much does it cost? :na

80. I'm home. :na

81. Dear, will call Tmorrow.pls accomodate. :na

82. First answer my question. :na

83. I only haf msn. It's yijue@hotmail.com :na

84. He is there. You call and meet him :na

85. Nah I don't think he goes to usf he lives around here though :na

86. Did you hear about the new Divorce Barbie It comes with all of Kenś stuff: na

87. SMS. ac Sptv: The New Jersey Devils and the Detroit Red Wings play Ice Hockey. Correct or Incorrect? End? Reply END SPTV :na

88. Congrats! 1 year special cinema pass for 2 is yours. call 09061209465 now! C Suprman V, Matrix3, StarWars3, etc all 4 FREE! bx420-ip4-5we. 150pm. Dont miss out! :na

6.9 Implications on the design of the phone application

In the book titled 'In Love and in Danger,' Levy describes ways to identify relationship abuse. He describes insecurities that abuse victims have, for example, "I thought things would change" and "I thought maybe it is me that is the problem." Both of these phrases show the self-blaming nature of some victims and speak to the fact that it is challenging to detect dating abuse if you are in an abusive relationship. The purpose of our machine learning algorithm is to identify abusive text messages to help users get an unbiased review of their digital communication [48]. Throughout the chapter, we describe the process of creating and validating an initial dataset, training different machine learning algorithms and evaluating the performance of these classifiers on three different variations of the unseen dataset. This forms the backbone for the phone application. As this application is targeted towards college-aged women, the training set was created to represent the socio-linguistic patterns of this demographic. As mentioned earlier, the text-based conversation usually involves spelling and grammatical errors and abbreviations. During the dataset creation process, we asked the participants to simulate these patterns as much as possible to get an accurate representation. A representative training set and a classifier with high accuracy will directly influence the effectiveness of this app and the general willingness to use this application. Six different evaluation parameters were used (Accuracy, Confusion matrix, F1-score and ROC AUC Score, Precision and Recall) to determine which classifier labels both abusive and non-abusive text messages accurately. The combination of linear support vector machine, unigram input, and tf-idf feature extractor had the best accuracy (91.6 %) with the unseen SPAM test dataset. The results of the confusion matrix illustrated that some of the classifiers performed well when labeling either abusive or non-abusive text messages, but we chose the most balanced classifier, i.e., the one that labeled non-abusive and abusive texts equally well. The combination of linear support vector machine, unigram input, and tf-idf labeled both these categories with balanced accuracy (Figure 6.4). This was consistent across all three datasets. For the F1-score, the same classifier and input combination outperformed other classifiers. For the design of this application, we are focusing on the balance of the classifier between abusive and non-abusive text messages. The goal is to avoid raising unnecessary false positive or false negatives while labeling the texts as accurately as possible. The results from the F1 score and ROC AUC score further support our choice of a balanced classifier, feature extractor, and n-gram input. Every relationship is unique, and people communicate differently. From the design standpoint, a 100% accuracy is not necessary. The users will be able to see the text messages that are labeled as abusive or non-abusive. Since the user makes the final judgment of a text being abusive or non-abusive, we hypothesize that some percentage of misclassification will not invalidate the usefulness of the application and deter users from trusting it.

6.10 Discussion

The final goal of this dissertation is to build a phone application to detect digital dating abuse from text messages. The work discussed in Chapter 5 is intended to support the backend of this application. The results of the annotation study and inter-rater agreement calculations using Light's Kappa indicated a moderate level of overall agreement among annotators when labeling the text messages as abusive and non-abusive. We can conclude that the dataset created after the labeling process can be considered as an abusive text message dataset. In the classification evaluation phase, we explored a combination of classifiers and feature extractors to identify the best possible combination that will give us the most accurate results. We conducted a 5-fold cross-validation and used a testing set to evaluate the accuracy of the classifiers. Although we have found no other published studies in this area of research, we are able to compare our accuracy with studies carried out in the broader domain related to interpersonal violence prevention and mental health. The accuracy value of eighty-seven point five percent (87.5%) is closely aligned with the performance evaluation data collected in the studies that were described in Section 5.1. The following chapter describes the prototype phone app with a working backend machine learning algorithm which is a combination of a linear support vector machine classifier, unigram input, and tf-idf feature extractor on the unseen test set.

Chapter 7

SecondLook- Development of the Prototype Android Phone Application

7.1 SecondLook: Features

Figure 7.1 shows above the phone app (SecondLook) consist of three features and several sub-features which are listed below.

- A. Understanding Digital Dating Abuse
- a. Healthy Relationships – The relationship spectrum
- b. FAQs
- c. Healthy Relationship Quiz
- B. Detection: Analyze Text Messages
- C. List of resources for help
- a. The results of the focus group study (as described in Chapter 3) influenced us to answer three questions a potential user has. Firstly, what is dating digital dating abuse and are there

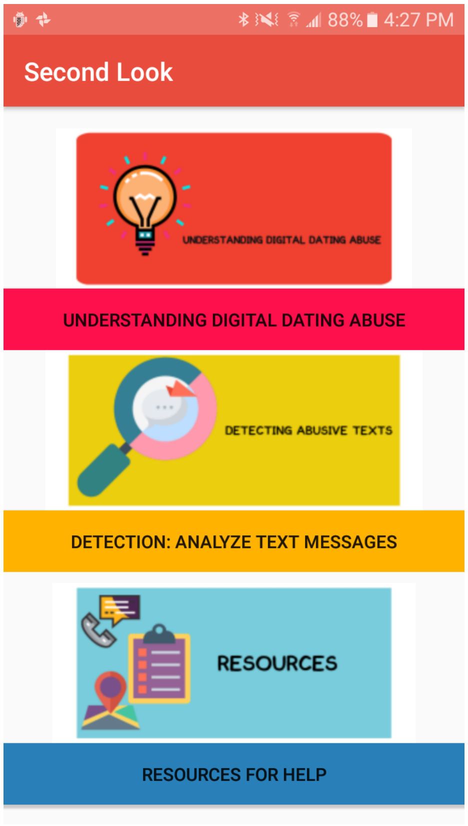

88% 4:27 PM
Second Look
UNDERSTANDING DIGITAL DATING ABUSE
UNDERSTANDING DIGITAL DATING ABUSE
DETECTING ABUSIVE TEXTS
DETECTION: ANALYZE TEXT MESSAGES
RESOURCES
RESOURCES FOR HELP

Figure 7.1: Illustrates the home screen of the phone app, which shows the three features (Feature A: Understanding Digital Dating Abuse, Feature B: Detection: Analyze Text Messages and Feature C: Resources for Help)

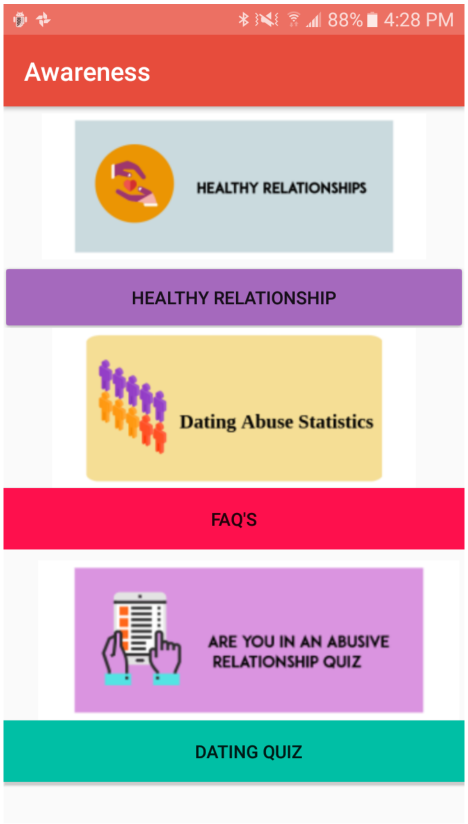
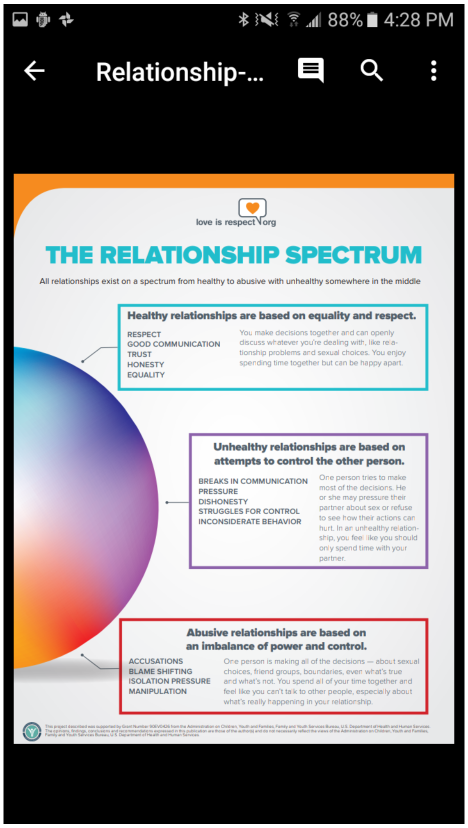


Figure 7.2: Illustrates the menu inside Feature A.

any data to prove this is a significant health concern, secondly how do I [user] know I [user] am in an abusive relationship and finally what are the resources that can be used for helping me [user] or others.

7.2 Feature A: Understanding Digital Dating Abuse

The content used to create Feature A has been developed and made publicly available by Loveisrespect which is a project of the National Domestic Violence Hotline funded through Administration on Children, Youth and Families, Family and Youth Services Bureau, U.S. Department of Health and Human Services [50]. These resources have not been modified; they are only presented



THE RELATIONSHIP SPECTRUM

All relationships exist on a spectrum from healthy to abusive with unhealthy somewhere in the middle

Healthy relationships are based on equality and respect.

RESPECT
GOOD COMMUNICATION
TRUST
HONESTY
EQUALITY

You make decisions together and can openly discuss whatever you're dealing with, like relationship problems and sexual choices. You enjoy spending time together but can be happy apart.

Unhealthy relationships are based on attempts to control the other person.

BREAKS IN COMMUNICATION
PRESSURE
DISHONESTY
STRUGGLES FOR CONTROL
INCONSIDERATE BEHAVIOR

One person tries to make most of the decisions. He or she may pressure their partner about sex or refuse to see how their actions can hurt. In an unhealthy relationship, you feel like you should only spend time with your partner.

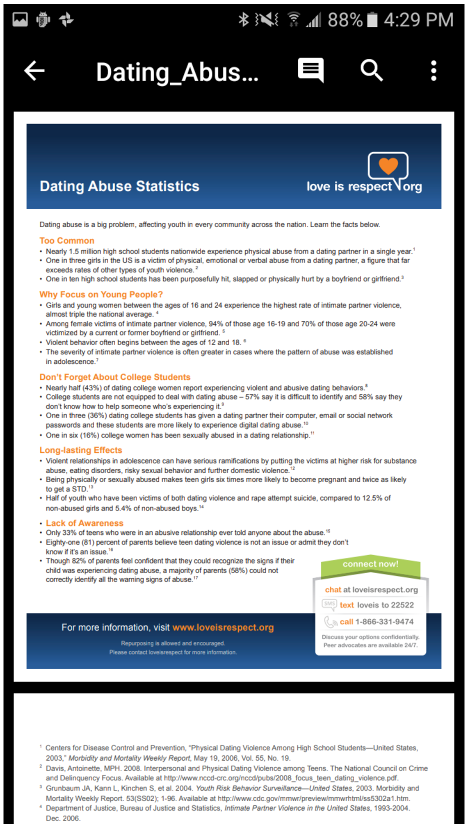
Abusive relationships are based on an imbalance of power and control.

ACCUSATIONS
BLAME SHIFTING
ISOLATION PRESSURE
MANIPULATION

One person is making all of the decisions — about sexual choices, friend groups, boundaries, even what's true and what's not. You spend all of your time together and feel like you can't talk to other people, especially about what's really happening in your relationship.

This project described was supported by Grant Number 90EV0426 from the Administration on Children, Youth and Families, Family and Youth Services Bureau, U.S. Department of Health and Human Services. The opinions, findings, conclusions and recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the Administration on Children, Youth and Families, Family and Youth Services Bureau, U.S. Department of Health and Human Services.

Figure 7.3: Shows the page with FAQs related to Dating Abuse Statistics.



Dating Abuse Statistics

Dating abuse is a big problem, affecting youth in every community across the nation. Learn the facts below.

Too Common

- Nearly 1.5 million high school students nationwide experience physical abuse from a dating partner in a single year.[1]
- One in three girls in the US is a victim of physical, emotional or verbal abuse from a dating partner, a figure that far exceeds rates of other types of youth violence.[2]
- One in ten high school students has been purposefully hit, slapped or physically hurt by a boyfriend or girlfriend.[3]

Why Focus on Young People?

- Girls and young women between the ages of 16 and 24 experience the highest rate of intimate partner violence, almost triple the national average.[4]
- Among female victims of intimate partner violence, 94% of those age 16-19 and 70% of those age 20-24 were victimized by a current or former boyfriend or girlfriend.[5]
- Violent behavior often begins between the ages of 12 and 18.[6]
- The severity of intimate partner violence is often greater in cases where the pattern of abuse was established in adolescence.[7]

Don't Forget About College Students

- Nearly half (43%) of dating college women report experiencing violent and abusive dating behaviors.[8]
- College students are not equipped to deal with dating abuse – 57% say it is difficult to identify and 58% say they don't know how to help someone who's experiencing it.[9]
- One in three (36%) dating college students has given a dating partner their computer, email or social network passwords and these students are more likely to experience digital dating abuse.[10]
- One in six (16%) college women has been sexually abused in a dating relationship.[11]

Long-lasting Effects

- Violent relationships in adolescence can have serious ramifications by putting the victims at higher risk for substance abuse, eating disorders, risky sexual behavior and further domestic violence.[12]
- Being physically or sexually abused makes teen girls six times more likely to become pregnant and twice as likely to get a STD.[13]
- Half of youth who have been victims of both dating violence and rape attempt suicide, compared to 12.5% of non-abused girls and 5.4% of non-abused boys.[14]

- Lack of Awareness
- Only 33% of teens who were in an abusive relationship ever told anyone about the abuse.[15]
- Eighty-one (81) percent of parents believe teen dating violence is not an issue or admit they don't know if it's an issue.[16]
- Though 82% of parents feel confident that they could recognize the signs if their child was experiencing dating abuse, a majority of parents (58%) could not correctly identify all the warning signs of abuse.[17]

connect now!
chat at loveisrespect.org
text loveis to 22522
call 1-866-331-9474
Discuss your options confidentially.
Peer advocates are available 24/7.

For more information, visit www.loveisrespect.org

Repurposing is allowed and encouraged.
Please contact loveisrespect for more information.

[1] Centers for Disease Control and Prevention, "Physical Dating Violence Among High School Students—United States, 2003," *Morbidity and Mortality Weekly Report*, May 19, 2006, Vol. 55, No. 19.
[2] Davis, Antoinette, MPH. 2008. Interpersonal and Physical Dating Violence among Teens. The National Council on Crime and Delinquency Focus. Available at http://www.nccd-crc.org/nccd/pubs/2008_focus_teen_dating_violence.pdf.
[3] Grunbaum JA, Kann L, Kinchen S, et al. 2004. *Youth Risk Behavior Surveillance—United States*, 2003. Morbidity and Mortality Weekly Report. 53(SS02); 1-96. Available at http://www.cdc.gov/mmwr/preview/mmwrhtml/ss5302a1.htm.
[4] Department of Justice, Bureau of Justice and Statistics, *Intimate Partner Violence in the United States*, 1993-2004. Dec. 2006.

Figure 7.4: Iillustrates The relationship spectrum screen in the app.

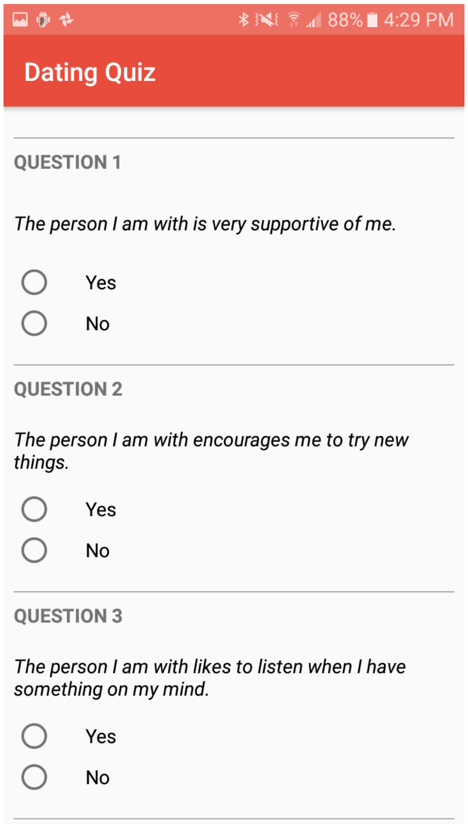
to the user in a mobile-friendly environment.

7.2.1 Healthy Relationships - The relationship spectrum

This feature is designed to help the user understand characteristics of healthy vs. unhealthy vs. abusive relationships. It provides scenarios for each kind of relationship and also words such as "Breakdown in communication" for unhealthy, "Equality" for a healthy relationship and "Manipulation" for abusive relationships to help users understand the relationship spectrum clearly. Figure 7.4 shows the page on the phone app [51].

7.2.2 Dating Abuse Statistics: FAQs

This feature contains a list of statistics related to Dating Abuse which has been collected and assimilated by various national and local organizations such as the Department of Justice, Bureau of



QUESTION 1

The person I am with is very supportive of me.

○ Yes
○ No

QUESTION 2

The person I am with encourages me to try new things.

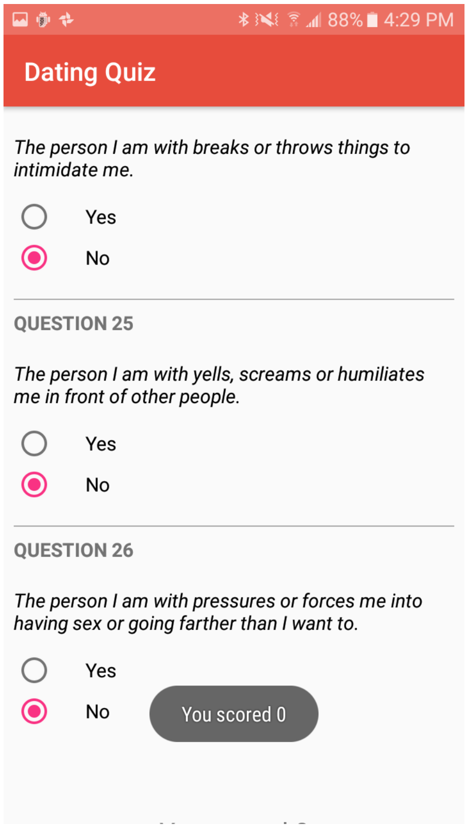
○ Yes
○ No

QUESTION 3

The person I am with likes to listen when I have something on my mind.

○ Yes
○ No

Figure 7.5: Illustrates the Dating Relationship Quiz as displayed on the app with each question and corresponding answer radio buttons.



The person I am with breaks or throws things to intimidate me.

○ Yes
● No

QUESTION 25

The person I am with yells, screams or humiliates me in front of other people.

○ Yes
● No

QUESTION 26

The person I am with pressures or forces me into having sex or going farther than I want to.

○ Yes
● No

You scored 0

Figure 7.6: Shows three questions that have been answered and a resulting score.

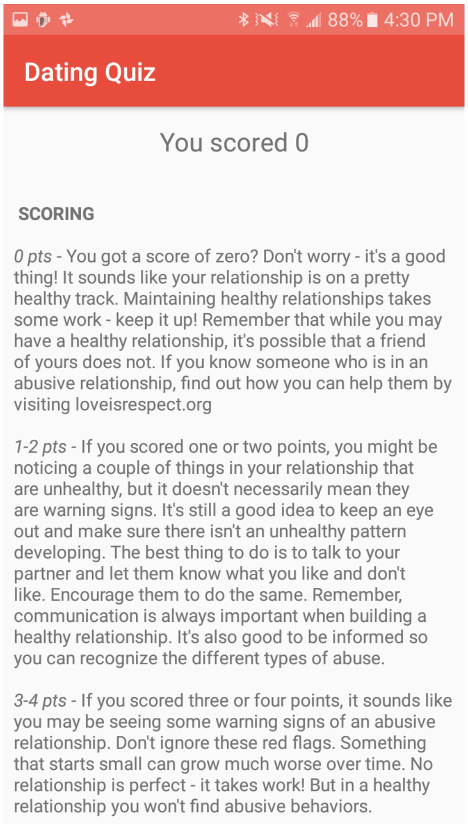
Justice Statistics and Centers for Disease Control and Prevention Refworks:899 . This aggregation of information lets the user know how rampant dating abuse is especially amongst girls and young women between the ages of 16 and 24 [19], school and college students, long-lasting effects of dating abuse and the lack of awareness. This feature is crucial for the phone app as it shows the relevance of dating abuse especially amongst young adults and college students (Figure 7.3).

7.2.3 Healthy Relationship Quiz

The Healthy Relationship Quiz feature is a twenty six-question yes and no questionnaire that is developed by Loveisrespect [5].This questionnaire is part of their resource category, which is publicly available for users online or offline. This questionnaire was modified to make it accessible for mobile phone users, and the users are provided with a score (Figure 7.5 and Figure 7.6). The scoring criteria, which is part of the resource package provided is adapted into the app's program. The user is also provided the scoring rubric, which allows them to see which category they fall under (Figure 7.7.). The scoring categories are zero, one-two points, three-four points and five points and above. Each category describes a particular type of relationship in conjunction with the relationship spectrum that was described in the feature "Healthy Relationships - The relationship spectrum."

7.3 Feature B- Detecting abusive text messages

The detection component of the phone application analyzes text messages that are exchanged between the user and their respective partner. For this application SMS (short messaging service), which is a text messaging service component of most mobile-device systems, is used to detect abusive text messages. SMS text messages were chosen as the input for several reasons: firstly, SMS texting is a widely used medium of communication all around the world especially amongst people in the United States. Although Internet-messaging services are getting increasingly popular, one of the reasons SMS texting is still popular as the top three American carriers have offered free SMS with almost all phone bundles since 2010, a stark contrast to Europe where SMS costs have been on the higher end [41]. In a study conducted by Pew all teens who spend time communicating with friends via different communication mediums. Text messaging, at 55%, is the most popular medium of communication [47]. Secondly, as digital dating abuse is a form of intimate partner violence, it is most unlikely that the perpetrator will use public social media posts to intimidate the victim. The user selects the text messages they want to be analyzed by choosing the contact number. They are provided three options; they can type the phone number associated with the contact or the contact name or select contact from the phone's contact list. Figures 7.8, 7.9 show the screens allowing the user to select the contact. The app then uses the selected contact's phone number to access all the text messages that were sent from this number to the user. These text messages are then uploaded to a server which uses a python based machine learning algorithm to classify the text messages as abusive or non-abusive. The processed data is then sent back to the phone app; Figure 7.10 illustrates the results page. Description related to the output is present at the top of the page to let the users understand the results better. Beside each text message, the corresponding label is displayed. A resources button is also located below the list of text messages, for quick access.



Dating Quiz

You scored 0

SCORING

0 pts - You got a score of zero? Don't worry - it's a good thing! It sounds like your relationship is on a pretty healthy track. Maintaining healthy relationships takes some work - keep it up! Remember that while you may have a healthy relationship, it's possible that a friend of yours does not. If you know someone who is in an abusive relationship, find out how you can help them by visiting loveisrespect.org

1-2 pts - If you scored one or two points, you might be noticing a couple of things in your relationship that are unhealthy, but it doesn't necessarily mean they are warning signs. It's still a good idea to keep an eye out and make sure there isn't an unhealthy pattern developing. The best thing to do is to talk to your partner and let them know what you like and don't like. Encourage them to do the same. Remember, communication is always important when building a healthy relationship. It's also good to be informed so you can recognize the different types of abuse.

3-4 pts - If you scored three or four points, it sounds like you may be seeing some warning signs of an abusive relationship. Don't ignore these red flags. Something that starts small can grow much worse over time. No relationship is perfect - it takes work! But in a healthy relationship you won't find abusive behaviors.

Figure 7.7: shows the screen which displays the user's score along with the rubric description.

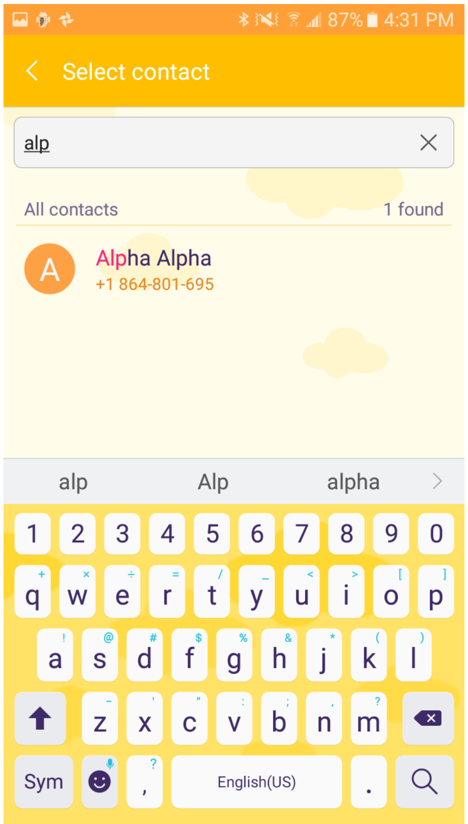


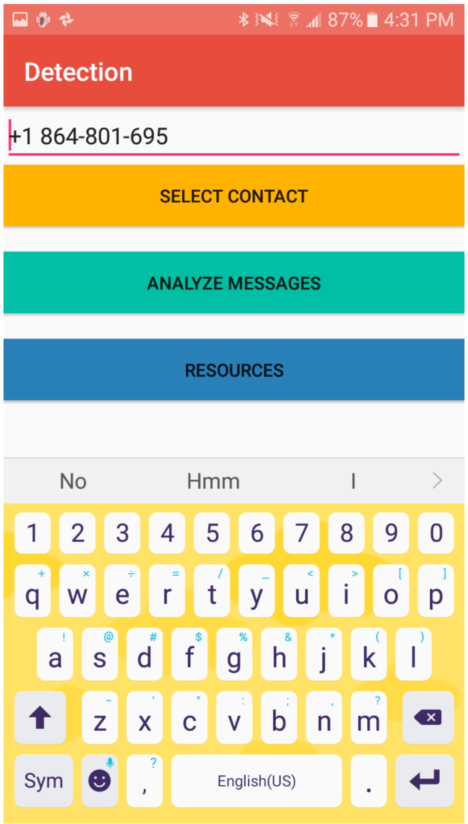
Figure 7.8: Shows the screens of the detection feature and different ways the user can select a contact(*type the contact name to select from the phone contacts*

7.4 FEATURE C Resource

The feature called resource (Fig. 7.11) contains a list of resources, such as contact information to the National Help Line, Counseling, and Psychological Services on campus, nearest support shelter, links to online resources and text message support. This aggregate of resources is provided to the user, to allow them to choose depending on what they think the severity of the situation is.

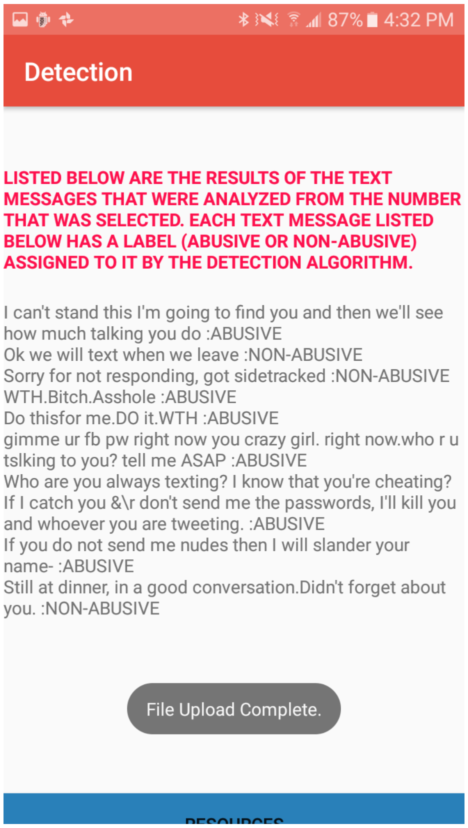
7.5 Software Requirements And Architecture

Android Studio 2.3.1 [26] was used to develop this phone app, after the initial user-interface being designed using a prototype designer called Balsamiq [83] . Both Features A (Understanding Digital Dating Abuse) and Feature C (Resources) are programmed using the Android studio IDE



+1 864-801-695

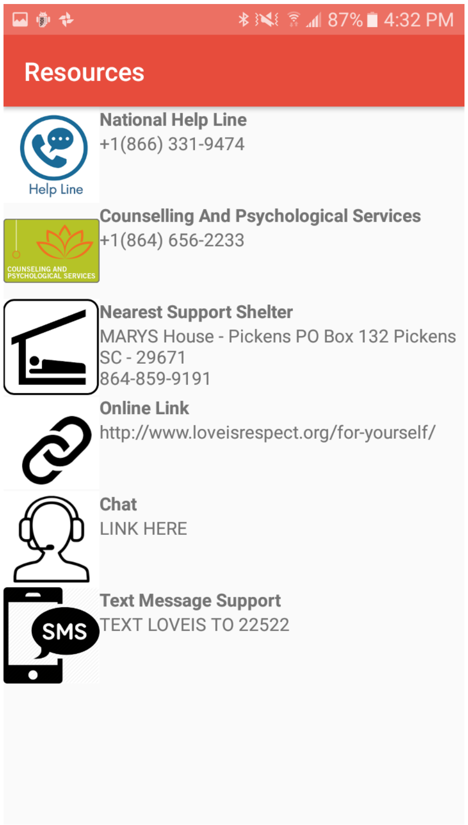
Figure 7.9: Shows the screens of the detection feature and different ways the user can select a contact *typing in the number manually*



LISTED BELOW ARE THE RESULTS OF THE TEXT MESSAGES THAT WERE ANALYZED FROM THE NUMBER THAT WAS SELECTED. EACH TEXT MESSAGE LISTED BELOW HAS A LABEL (ABUSIVE OR NON-ABUSIVE) ASSIGNED TO IT BY THE DETECTION ALGORITHM.

I can't stand this I'm going to find you and then we'll see how much talking you do :ABUSIVE
Ok we will text when we leave :NON-ABUSIVE
Sorry for not responding, got sidetracked :NON-ABUSIVE
WTH.Bitch.Asshole :ABUSIVE
Do thisfor me.DO it.WTH :ABUSIVE
gimme ur fb pw right now you crazy girl. right now.who r u tslking to you? tell me ASAP :ABUSIVE
Who are you always texting? I know that you're cheating? If I catch you &\r don't send me the passwords, I'll kill you and whoever you are tweeting. :ABUSIVE
If you do not send me nudes then I will slander your name- :ABUSIVE
Still at dinner, in a good conversation.Didn't forget about you. :NON-ABUSIVE

Figure 7.10: shows the phone screen witht analyzed text messages



87% 4:32 PM

Resources

National Help Line
+1(866) 331-9474

Counselling And Psychological Services
+1(864) 656-2233

Nearest Support Shelter
MARYS House - Pickens PO Box 132 Pickens SC - 29671
864-859-9191

Online Link
http://www.loveisrespect.org/for-yourself/

Chat
LINK HERE

Text Message Support
TEXT LOVEIS TO 22522

Figure 7.11: Illustrates the home screen of the phone app, which shows the three features (Feature A: Understanding Digital Dating Abuse, Feature B: Detection: Analyze Text Messages and Feature C: Resources for Help)

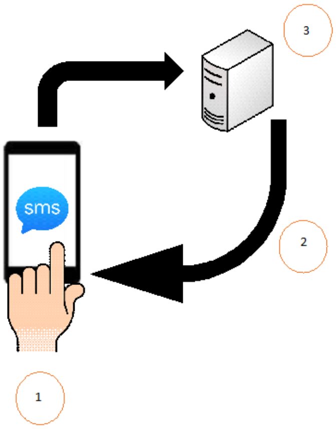


Figure 7.12: Illustrates the menu inside Feature A.

and completed in stage 1 as shown in Figure 7.12. The content of Feature A is stored as pdf files on a Google drive account and can be accessed by the user. The scores for the 'Healthy Relationship Quiz' are calculated by an Android based Java program within stage 1 and the results are displayed to the user. For Feature B both stage 2 and 3 are required for the detection application to work. In stage 2 (which is a python-based server called Flask) we host the machine learning algorithm (combination of linear support vector machine with unigram and tf-idf). The content uploaded from Stage 1 is processed in Stage 2 and in Stage 3 the server communicates with the android app to display the results of Stage 2 to the user on the phone screen as shown in Figure 7.11.

7.5.1 Functionalities of Stage 2 and Stage 3.

On selection of the contact on the Detection screen, a method searches for the SMSes from that particular contact in the Android SMS inbox. After it passes specific validation checks (example: there are messages from that contact), a new file is created in the primary external storage space of the Android device, and the SMSes are copied in understandable text format. Appropriate messages are displayed on the Android device screen after each stage, providing feedback to the user regarding the success or failure of each step (example: failure message if there is an exception while writing to the file). With the help of HTTP Post request method, data from the Android application is uploaded to the server. The HTTP connection is kept alive/open for the result data which comes from the server to the Android application after the sent file which contained the SMSes has been processed by the machine learning system in Python. The result data which comprises of the individual messages with appropriate abusive or non-abusive tags are now displayed on the device screen.

Chapter 8

Evaluating The usability of the Phone App and Understanding User Trust and Effectiveness

8.1 Research Questions

The previous chapter describes the software development of the Android prototype. In this chapter, we will be discussing the app's level of usability as evaluated by qualitative and quantitative usability metrics. The research questions we focused on for this study are:

1. Are the individual features of the app, i.e., Awareness, Detection, and Resources designed in an intuitive manner that helps the user acquire the information they require?

2. Presented with a set of text messages which contains both abusive and non-abusive text messages

   (a) What is the threshold of abusive text messages (in percentage or count) to be considered in an abusive relationship?

   (b) Does the way the data is visualized have an impact on user's trust when it comes to using the detection application?

Gender	
Female	77.78%
Male	22.22%
Ethnicity	
Caucasian	27.78%
Asian	38.89%
African American	16.67%
Hispanic	5.56%
Other	11.11%
Education	
Bachelor's Degree	33.33%
Doctorate Degree	16.67%
Master's Degree	50.00%
Marital Status	
In a relationship	22.22%
Married	44.44%
Single	33.33%

Table 8.1: Demographic Information of all the participants

(c) Is there a change in the user's perception of abusive vs. non-abusive in observing the text messages classified by the machine learning algorithm (as displayed on the app) vs. labeling the text messages themselves?

(d) What is the threshold of abusive text messages that encourages users to seek help and if they do, which is their most preferred resources?

8.2 Study Design and participants

We designed a within-subject study where every participant would be interacting with the phone app, complete a set of three tasks and provide qualitative and quantitative feedback. We recruited eighteen (n=18) participants to complete this study in a lab experiment setting. The participants were asked to complete a demographic questionnaire (Table 8.1), and a dating abuse awareness questionnaire. To get an accurate sample of our target population (college-aged women), we primarily recruited women from Clemson University and members of the community. Five out of all the participants were considered as expert participants as all have more than a year of experience conducting usability studies or doing user-experience research.

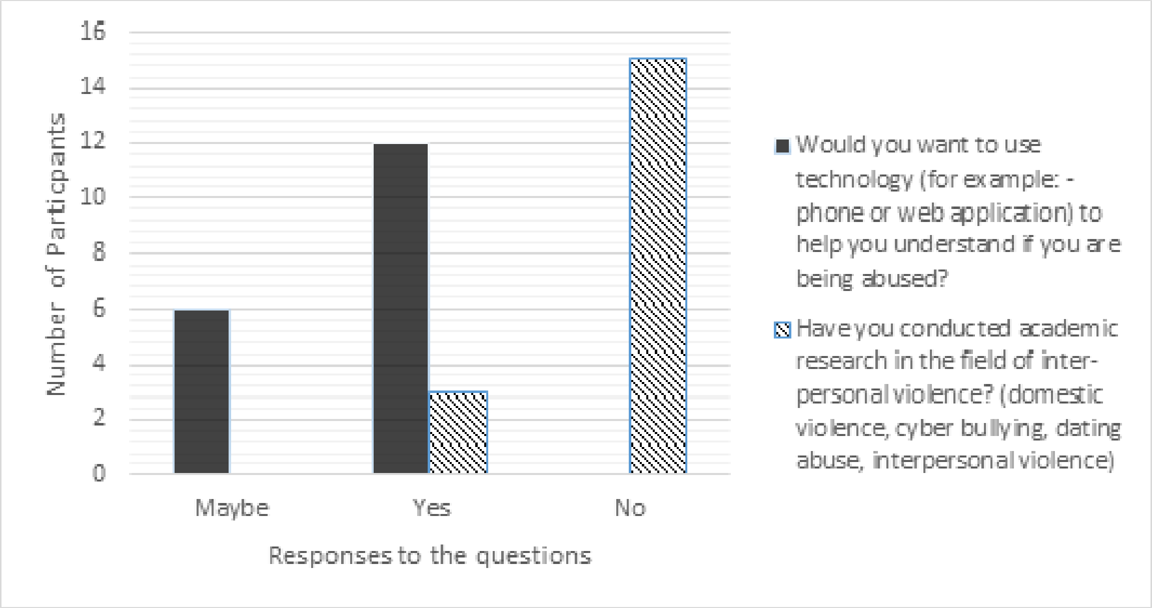


Figure 8.1: Shows the participants responses to the questions a) Have you conducted academic research in the field of interpersonal violence? (domestic violence, cyber bullying, dating abuse, interpersonal violence) b) Would you want to use technology (for example - phone or web application) to help you understand if you are being abused?

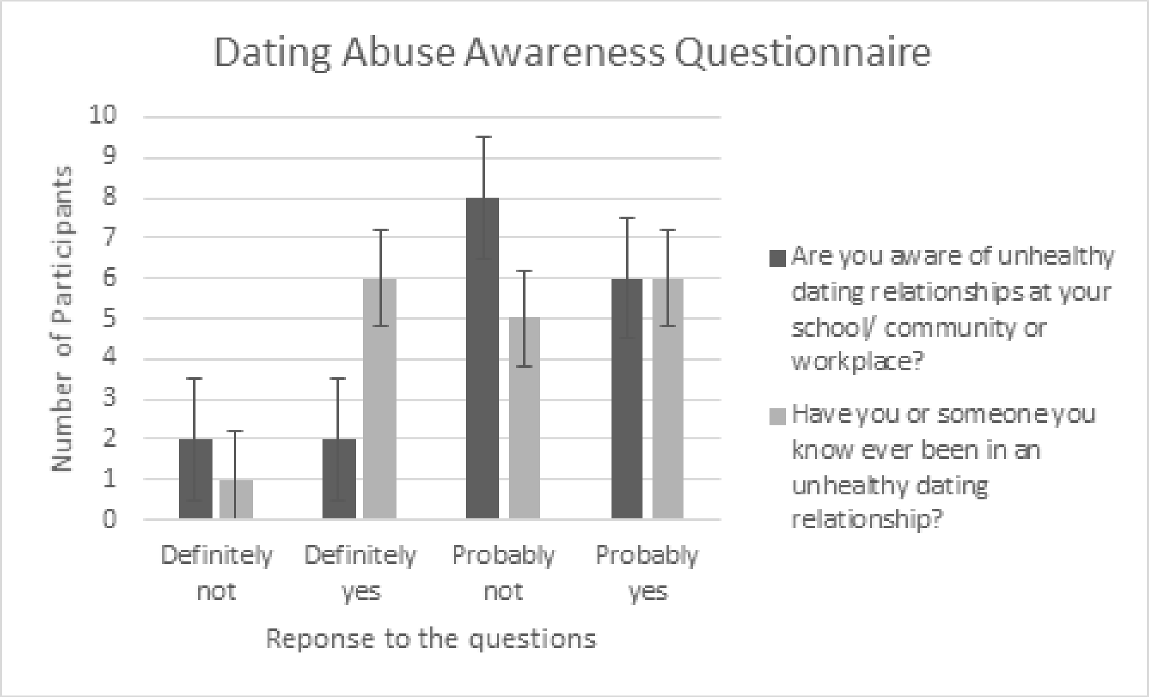


Figure 8.2: Shows the participants responses to the questions a) Are you aware of unhealthy dating relationships at your community or workplace? b) Have you or someone you know ever been in an unhealthy dating relationship?

8.3 Method

The participants were provided an Android phone that had the phone app installed, and the list of tasks was provided; they completed each task and answered the questions before completing the next one. While they completed the tasks, the participants were asked to verbally report about the process/opinions about the results and feedback on the design. Task 1

1. Use the feature "Understanding Digital Dating Abuse."
2. Click on the button "Understanding Digital Dating Abuse" and find out more about "Healthy Relationships."
3. Click on the button FAQs to know more about 'digital dating abuse.'
4. Complete the dating quiz.

Task 2

1. Use the feature "Detection" to select any contact number and choose the analyze feature to view the results.

Task 3

1. Select the "Resources" feature and observe the options provided to you and answer the following questions.

8.3.1 Measures

To evaluate each task we used a modified Heuristic evaluation Metric [55, 56] , heuristic evaluation is a method for finding the usability problems in a user interface design so that they can be attended to as part of an iterative design process . Table 8.2 and Table 8.3 list the questions that the users were asked after every task was completed and the severity rating scale.

Each task also included specific questions such as

1. After completing Task 1, the participants were asked to report the score they received after completing the healthy relationship quiz, what the score translated to and one characteristic of a healthy relationship. The rationale behind asking these three questions was to ascertain if the user had completed the quiz, read through the " Relationship Spectrum" and if they were able to locate the rubric related to each score and understood it clearly.

Visibility of System Status Are you kept aware about the system progress with appropriate feedback within reasonable time?
Match between System and Real World – Does the system use concepts and language familiar to the user rather than system-oriented terms.
Does the system use real-world conventions and display information in a natural and logical order?
User control - Can the users do what they want to do when they want to do?
Consistency and Standards -Do design elements such as objects and actions have the same meaning or effect in different situations? i.e.user knowledge required to use the system by letting users generalize from existing experience of the system or other systems.
Recognition rather than recall –Are design elements such as objects actions and options visible? Is the user forced to remember information from one part of a system to another?
Flexibility and efficiency of use – Are the task methods efficient and can users customize frequent actions or use shortcuts?
Aesthetic and minimalist design. Do dialogues contain irrelevant or rarely needed information? Dialogues should not contain information that is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility

Table 8.2: Modified heuristic evaluation metrics

2. After completing Task 2, the participants were asked

   (a) Based on the results presented, would you consider yourself in an abusive dating relationship?

   (b) Have the results been presented in an effective manner such that you understand the analysis of the text messages clearly?

   (c) Would you trust the information about the abusive text messages that the app has provided to you?

These three questions were asked to answer Research Question 2a and 2b. We used a 5-point Likert Scale (Strongly Agree, Somewhat Agree, Neither Agree or Disagree, Somewhat Disagree and Strongly Disagree) for measuring their responses [8]. We also asked the participants to read

0= I don't agree that this is a usability problem at all
1 = Cosmetic problem only need not be fixed unless extra time is available on the project
2 = Minor usability problem fixing this should be given low priority
3 = Major usability problem important to fix, so should be given high priority
4 = Usability catastrophe imperative to fix this before product can be released

Table 8.3: The scale used was a five-point severity scale where:

the text messages (which were identical to the ones displayed to them on the app) (FIGURE NUMBER) and based on the definition of digital dating abuse provided earlier, arrange the text messages into three categories (Abusive, Non-Abusive Not-Sure). The last question was asked to explore whether there is a change in the user's perception of abusive vs. non-abusive in observing the text messages classified by the machine learning algorithm (as displayed on the app) vs. labeling the text messages themselves?

3. After completing Task 3, the participant was asked 1) Based on the text message analysis that was shown to you in the detection feature, would you get help? They were also asked to rank each resource in the order of preference (they were instructed that 1 would indicate their primary choice).

8.4 Results

8.4.1 Task 1

Table 8.4, 8.5 and 8.6 illustrates the results from Task1, 2 and 3 respectively, the first column contains all the questions, and every column is a participant's response to the questions. For Task 1, for the first four participants, only the button Understanding Digital Dating Abuse, Detection Analyze Text messages, and Resources for help (circled in the green in the image below in Figure 8.2) was clickable. All four participants clicked on the images above the buttons (circled in pink in Figure 8.3) in the figure, and expressed their confusion regarding how they would proceed to the next task. As this was a recurrent issue for the next screen too, for the next fourteen participants, we modified the application in such a way, both the images and buttons were clickable and led to the same page. This considerably decreased participant frustrations but some participants stressed

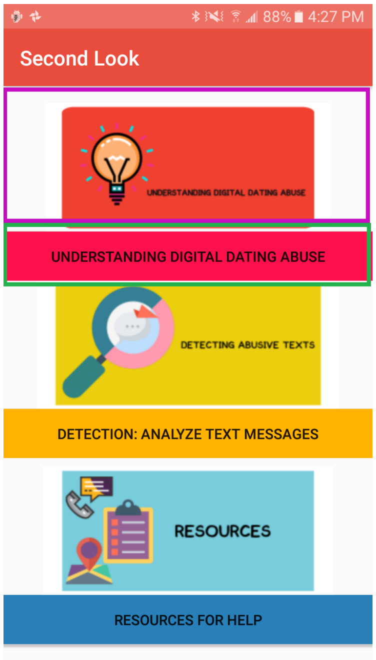
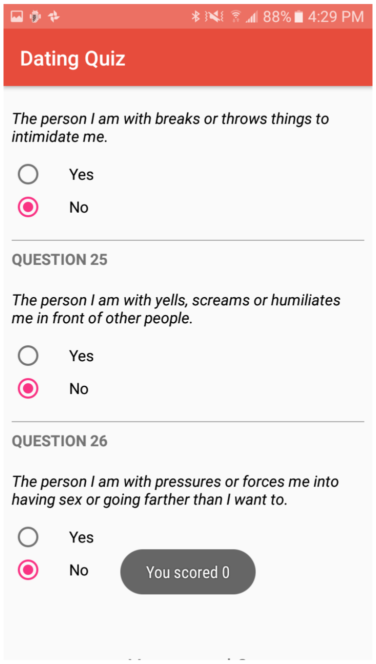

88% 4:27 PM
Second Look
UNDERSTANDING DIGITAL DATING ABUSE
UNDERSTANDING DIGITAL DATING ABUSE
DETECTING ABUSIVE TEXTS
DETECTION: ANALYZE TEXT MESSAGES
RESOURCES
RESOURCES FOR HELP

Figure 8.3: Home Page



The person I am with breaks or throws things to intimidate me.

○ Yes
● No

QUESTION 25

The person I am with yells, screams or humiliates me in front of other people.

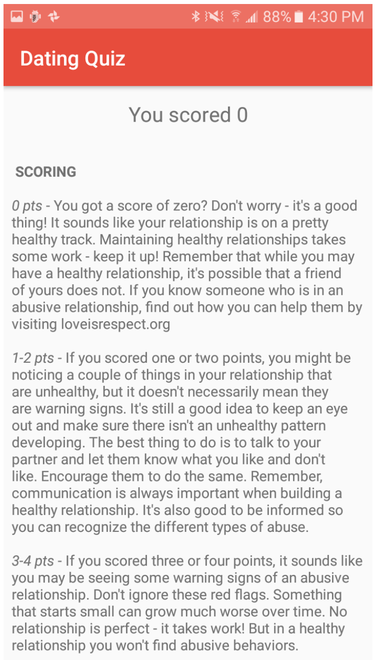
○ Yes
● No

QUESTION 26

The person I am with pressures or forces me into having sex or going farther than I want to.

○ Yes
● No

Figure 8.4: Quiz Page with Score



You scored 0

SCORING

0 pts - You got a score of zero? Don't worry - it's a good thing! It sounds like your relationship is on a pretty healthy track. Maintaining healthy relationships takes some work - keep it up! Remember that while you may have a healthy relationship, it's possible that a friend of yours does not. If you know someone who is in an abusive relationship, find out how you can help them by visiting loveisrespect.org

1-2 pts - If you scored one or two points, you might be noticing a couple of things in your relationship that are unhealthy, but it doesn't necessarily mean they are warning signs. It's still a good idea to keep an eye out and make sure there isn't an unhealthy pattern developing. The best thing to do is to talk to your partner and let them know what you like and don't like. Encourage them to do the same. Remember, communication is always important when building a healthy relationship. It's also good to be informed so you can recognize the different types of abuse.

3-4 pts - If you scored three or four points, it sounds like you may be seeing some warning signs of an abusive relationship. Don't ignore these red flags. Something that starts small can grow much worse over time. No relationship is perfect - it takes work! But in a healthy relationship you won't find abusive behaviors.

Figure 8.5: Quiz Page with rubric

the point that this was a functional fix but a design redundancy.

When completing Task 1a and 1b, i.e., reading about Healthy Relationships and FAQs, one of the participants (expert) said that the pdf documents we were using for the app, “was not optimized for a mobile phone application.” Other participants were able to use the inbuilt feature in the phone such as finder tap zooming and changing the screen orientation to read the material intended. While completing Task 1c) the participants were able to complete the Healthy Relationship Quiz, but after they saw the score, some participants were alarmed at the score they received, as it only displayed the score and not the rubric (Figure 8.4). To view the rubric the participants had to scroll down, but there was no visual cue that was guiding them to do that. One participant (expert) mentioned, “There was no feedback to say that I completed the task or any end trail.”

8.4.2 Task 2

For Task 2, the HEM scores are listed in Table 8.5. The scores and verbal feedback received from the first four participants (2 experts) all listed two primary issues with the user-interface design. Firstly, the application was designed so that the Analyze button appears on the screen if the contact is selected which would result in a tap on the select contact box. This event triggered the Analyze

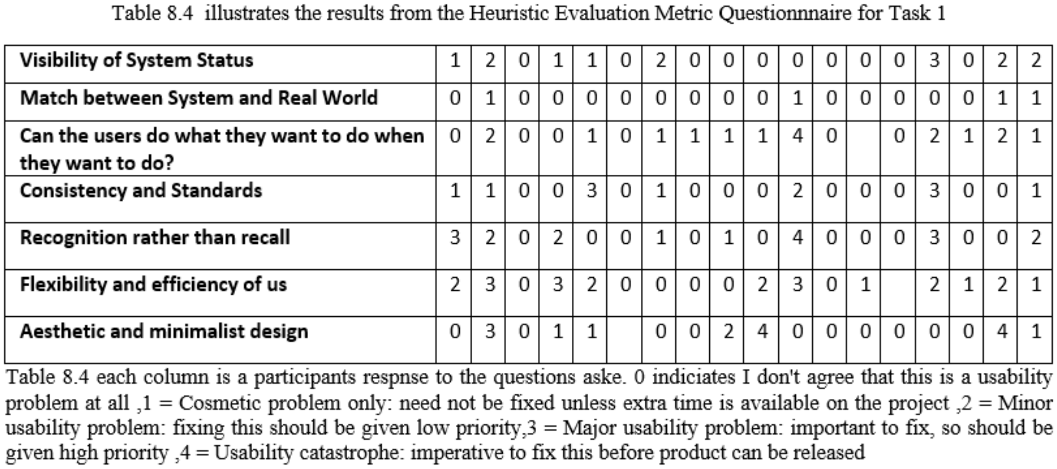


Table 8.4 illustrates the results from the Heuristic Evaluation Metric Questionnnaire for Task 1

Visibility of System Status	1	2	0	1	1	0	2	0	0	0	0	0	0	0	3	0	2	2
Match between System and Real World	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1
Can the users do what they want to do when they want to do?	0	2	0	0	1	0	1	1	1	1	4	0		0	2	1	2	1
Consistency and Standards	1	1	0	0	3	0	1	0	0	0	2	0	0	0	3	0	0	1
Recognition rather than recall	3	2	0	2	0	0	1	0	1	0	4	0	0	0	3	0	0	2
Flexibility and efficiency of us	2	3	0	3	2	0	0	0	0	2	3	0	1		2	1	2	1
Aesthetic and minimalist design	0	3	0	1	1		0	0	2	4	0	0	0	0	0	0	4	1

Table 8.4 each column is a participants respnse to the questions aske. 0 indiciates I don't agree that this is a usability problem at all ,1 = Cosmetic problem only: need not be fixed unless extra time is available on the project ,2 = Minor usability problem: fixing this should be given low priority,3 = Major usability problem: important to fix, so should be given high priority ,4 = Usability catastrophe: imperative to fix this before product can be released

Figure 8.6: Illustrates the results from the Heuristic Evaluation Metric Questionnaire for Task 1

button to appear on the screen. But none of the first fours participants followed the workflow that was designed. This lead to comments such as " The select button did not change, how I am supposed to know I am supposed to click on it," "Where is the Analyze button? Do I need to type something else?" and on average participants spent 40 seconds to find the button. To reduce user frustration and confusion, we changed the design flow of the app such that the Analyze Button would appear on the screen along with the "Select the Contact" button. For the next fourteen participants, we tested the modified user-interface and received no negative feedback related to this particular issue.

For the first four participants, when the results from the classifier were displayed (Figure 8.5) the message (as circled in the pink box) was not displayed. The participants were unable to understand what these results meant, and comments such as " Where are the labels?" was made. We modified the user-interface for the following fourteen participants, and the message as shown in Figure 8.6 was included for the participants. Participants on an average spent 120 secs on this page to read the text messages. Majority of the participants complained about the display not being intuitive and suggested ways to make the text more distinguishable such as one participant suggested " Separate the text messages by color" and "Put it in a table with columns to make it clear", "It is difficult to read the text messages, make the text larger and color code the labels". The Heuristic Evaluation Metrics reveal the usability issue that the participants faced and suggested

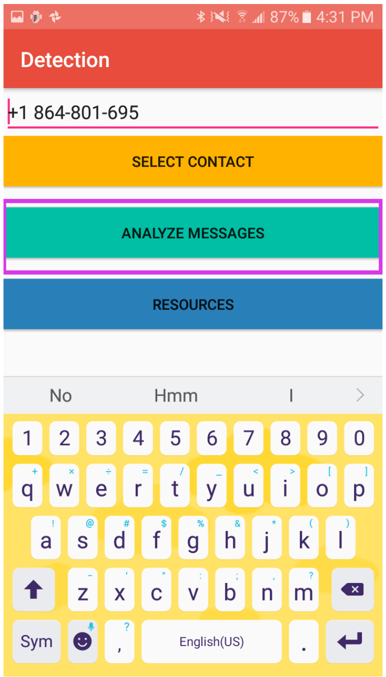
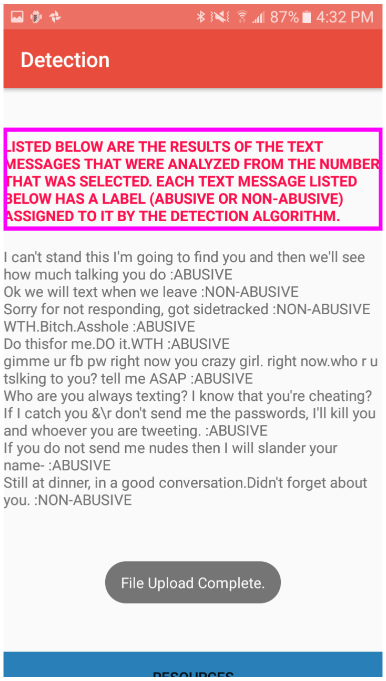


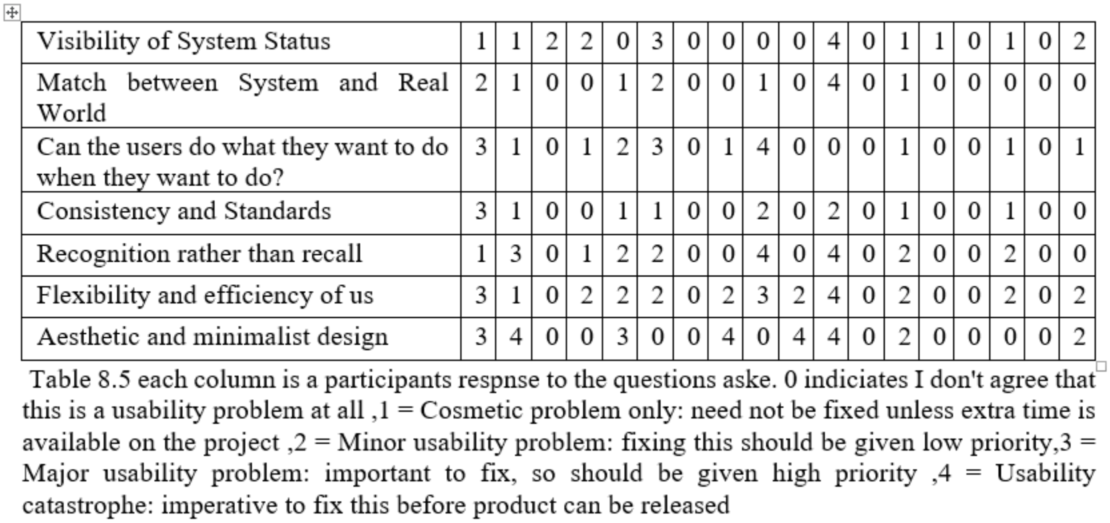
Figure 8.7: The pink boxes indicate features that were added after feedback was received from four participants.



LISTED BELOW ARE THE RESULTS OF THE TEXT MESSAGES THAT WERE ANALYZED FROM THE NUMBER THAT WAS SELECTED. EACH TEXT MESSAGE LISTED BELOW HAS A LABEL (ABUSIVE OR NON-ABUSIVE) ASSIGNED TO IT BY THE DETECTION ALGORITHM.

```
I can't stand this I'm going to find you and then we'll see
how much talking you do :ABUSIVE
Ok we will text when we leave :NON-ABUSIVE
Sorry for not responding, got sidetracked :NON-ABUSIVE
WTH.Bitch.Asshole :ABUSIVE
Do thisfor me.DO it.WTH :ABUSIVE
gimme ur fb pw right now you crazy girl. right now.who r u
tslking to you? tell me ASAP :ABUSIVE
Who are you always texting? I know that you're cheating?
If I catch you &\r don't send me the passwords, I'll kill you
and whoever you are tweeting. :ABUSIVE
If you do not send me nudes then I will slander your
name- :ABUSIVE
Still at dinner, in a good conversation.Didn't forget about
you. :NON-ABUSIVE
```

Figure 8.8: The pink boxes indicate features that were added after feedback was received from four participants



Visibility of System Status	1	1	2	2	0	3	0	0	0	0	4	0	1	1	0	1	0	2
Match between System and Real World	2	1	0	0	1	2	0	0	1	0	4	0	1	0	0	0	0	0
Can the users do what they want to do when they want to do?	3	1	0	1	2	3	0	1	4	0	0	0	1	0	0	1	0	1
Consistency and Standards	3	1	0	0	1	1	0	0	2	0	2	0	1	0	0	1	0	0
Recognition rather than recall	1	3	0	1	2	2	0	0	4	0	4	0	2	0	0	2	0	0
Flexibility and efficiency of us	3	1	0	2	2	2	0	2	3	2	4	0	2	0	0	2	0	2
Aesthetic and minimalist design	3	4	0	0	3	0	0	4	0	4	4	0	2	0	0	0	0	2

Table 8.5 each column is a participants respnse to the questions aske. 0 indiciates I don't agree that this is a usability problem at all ,1 = Cosmetic problem only: need not be fixed unless extra time is available on the project ,2 = Minor usability problem: fixing this should be given low priority,3 = Major usability problem: important to fix, so should be given high priority ,4 = Usability catastrophe: imperative to fix this before product can be released

Figure 8.9: Illustrates the results from the Heuristic Evaluation Metric Questionnaire for Task 2

modifications.

8.4.3 Task 3

For Task 3, most participants found some major usability issues and the scores for the questions “Can the users do what they want to do when they want to do” and “Flexibility and efficiency of use” show us that. The major feedback we received was that they are unable to click on the icons which displayed contact information or website links. These were static images with texts and hence did not provide any visual or textual feedback. The participants found this aspect of the interface limiting and hence not flexible for use. Some comments our expert participants mentioned were “ I don’t want to type the phone number in if I want to call the helpline in. It should let me call right away” and “ The link is not working [while clicking on a webpage link].”

Along with the heuristic evaluation metrics for Task 2, we asked the user three questions which they answered on 5 point Likert Rating Scale. Figure 10 shows the responses to the three questions 1) Based on the results presented would you consider yourself in an abusive dating relationship? - Twelve out of the eighteen participants (66.67 %) reported they strongly disagreed. For the question 2) Have the results been presented in an effective manner such that you understand the analysis of the text messages clearly? - 50% of the participants Strongly Agreed and 33.33% somewhat agreed with that statement. For the question 3) Would you trust the information about

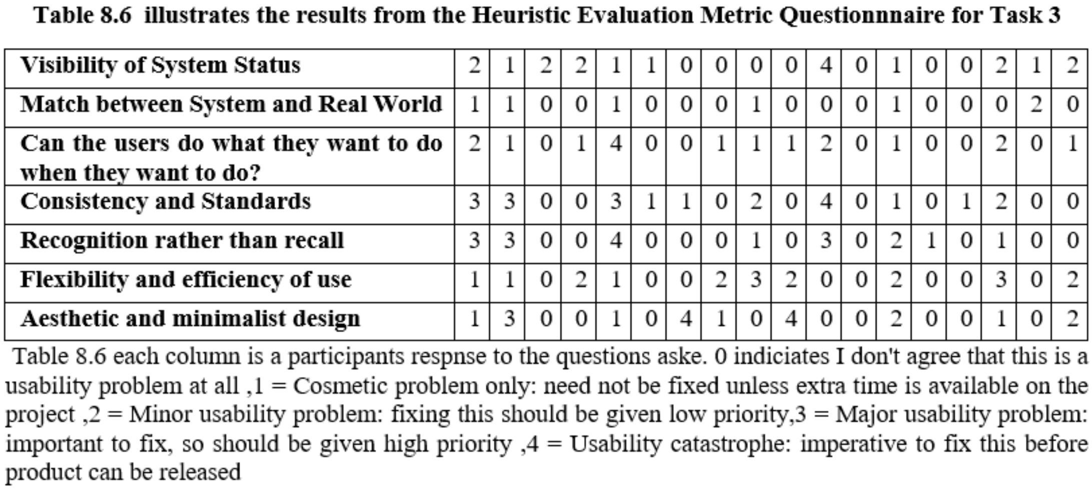


Table 8.6 illustrates the results from the Heuristic Evaluation Metric Questionnaire for Task 3

Visibility of System Status	2	1	2	2	1	1	0	0	0	0	4	0	1	0	0	2	1	2
Match between System and Real World	1	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	2	0
Can the users do what they want to do when they want to do?	2	1	0	1	4	0	0	1	1	1	2	0	1	0	0	2	0	1
Consistency and Standards	3	3	0	0	3	1	1	0	2	0	4	0	1	0	1	2	0	0
Recognition rather than recall	3	3	0	0	4	0	0	0	1	0	3	0	2	1	0	1	0	0
Flexibility and efficiency of use	1	1	0	2	1	0	0	2	3	2	0	0	2	0	0	3	0	2
Aesthetic and minimalist design	1	3	0	0	1	0	4	1	0	4	0	0	2	0	0	1	0	2

Table 8.6 each column is a participants respnse to the questions aske. 0 indiciates I don't agree that this is a usability problem at all ,1 = Cosmetic problem only: need not be fixed unless extra time is available on the project ,2 = Minor usability problem: fixing this should be given low priority,3 = Major usability problem: important to fix, so should be given high priority ,4 = Usability catastrophe: imperative to fix this before product can be released

Figure 8.10: Illustrates the results from the Heuristic Evaluation Metric Questionnaire for Task 3

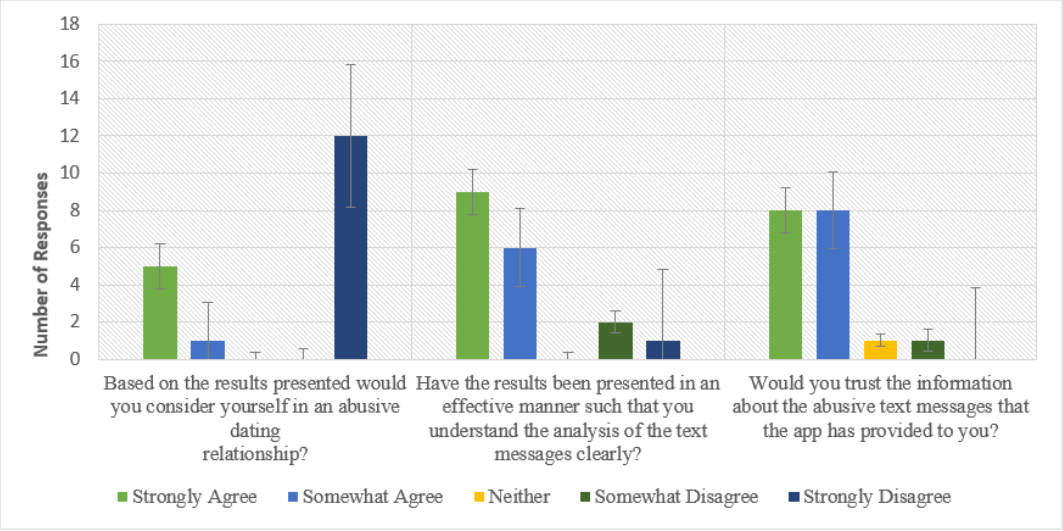


Figure 8.10 shows the participants responses to the questions 1) *Based on the results presented would you consider yourself in an abusive dating relationship?* 2) *Have the results been presented in an effective manner such that you understand the analysis of the text messages clearly?* 3) *Would you trust the information about the abusive text messages that the app has provided to you?*

Figure 8.11: Participants responses to the following questions

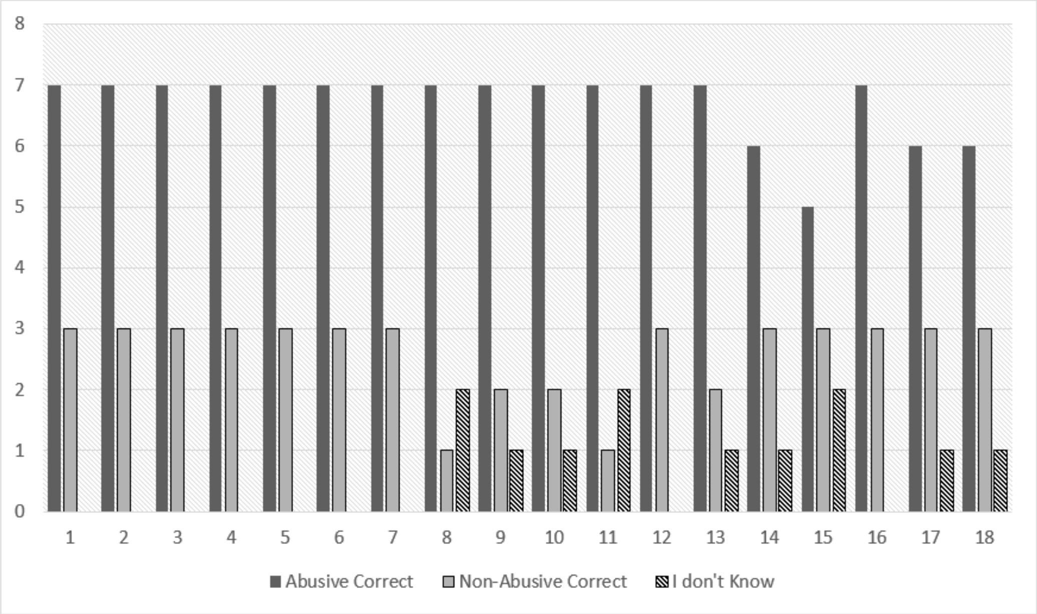


Figure 8.11 shows the number of correct abusive vs non-abusive vs I don't know responses provided by each participants (x axis shows each participant id).

Figure 8.12: User labelled correct abusive text messages

the abusive text messages that the app has provided to you? Eight of the participants Strongly Agreed and Eight Somewhat Agreed (44.44% each).

The participants were also asked to label the list of text messages that they viewed on the screen as abusive, non-abusive and don't know. Figure 8.11 shows the results of the responses. They list provided consisted of 7 abusive and 3 non-abusive text messages which had been validated by the domain experts as mentioned in Chapter 5. Fourteen out of eighteen participants labeled the abusive text messages correctly, and nine out of eighteen participants labeled both abusive and non-abusive text messages correctly. The text message "One of my friends said he saw you walking on campus? I thought you were busy all day today. Shouldn't you be in the library? If you had all this free time, you should've been it at my place" was the misclassified text message (either non-abusive or I don't know). For Task 3, the participants were asked the following questions "based on the text message analysis that was shown to you in the detection feature, would you get help?" Eleven out of eighteen participants strongly agreed to get help (Figure 8.12). The participants were also asked to rank the list of resources in order of preference (1 being the most preferred), Table 1 shows the ranks the participants assigned for every resource.

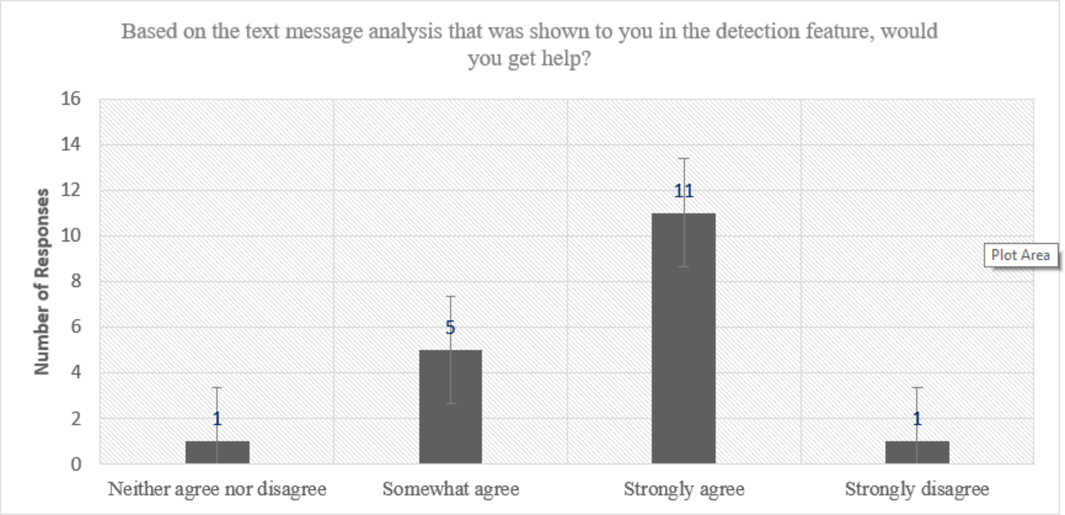
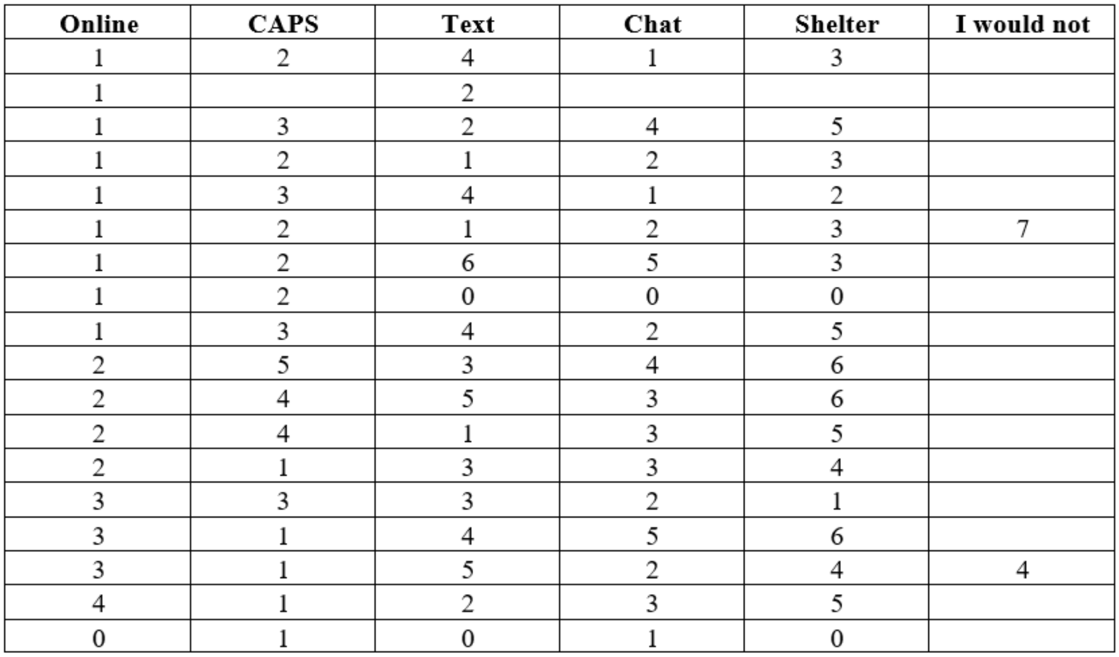


Figure 8.13: Responses to questions: Based on the text message analysis that was shown to you in the detection feature, would you get help?



Online	CAPS	Text	Chat	Shelter	I would not
1	2	4	1	3	
1		2			
1	3	2	4	5	
1	2	1	2	3	
1	3	4	1	2	
1	2	1	2	3	7
1	2	6	5	3	
1	2	0	0	0	
1	3	4	2	5	
2	5	3	4	6	
2	4	5	3	6	
2	4	1	3	5	
2	1	3	3	4	
3	3	3	2	1	
3	1	4	5	6	
3	1	5	2	4	4
4	1	2	3	5	
0	1	0	1	0	

Table 8.7 – Shows the rank each participant provided to the list of resources. 1 indicates the one that would be used most frequently.

Figure 8.14: Shows the rank each participant provided to the list of resources. 1 indicates the one that would be used most frequently.

From the results, Online resources such as (a link to loveisnotabuse.org) were the highest rated resource, followed by chat support, counseling and psychological services, text message support and women's shelter. Participants also had the option to suggest an alternate resource; participants suggested "Talking to friends, supported family members" and one participant suggested "Talk to the sender about the tone of messages being sent to me" and ranked it as the primary resource.

8.5 Discussion

From the user-interface design standpoint, the results of both quantitative and qualitative feedback revealed specific major usability issues. Such as for the "Healthy Relationship Quiz," the lack of a visual cue to inform the users about scrolling down to view the rubric, the labeled text messages in the Detection page are not easily readable and need to be structured more intuitively. Another major concern that all the participants expressed was the lack of a back button. While designing the app, we designed it to use Android's inbuilt features such as the back button on the phone to be used as the default backward navigation feature. But the participants expressed their concern about not being able to understand that navigation feature and suggested we build in a navigation system within the app's user interface such as including a shortcut on each screen which allows the user's flexibility to navigate the app without having to return to the home screen multiple times. All these suggestions will be incorporated into future iterations of the phone app and will be discussed in detail in the Future Work section.

The participant responses also concluded that the text messages were labeled accurately and over 50% of the participants strongly agreed, and 33.33% participants somewhat agreed that the data was effectively displayed and about 44.44% participants strongly agreed or somewhat agreed that the results where trustworthy. These results strongly indicate that the detection application was effective and trustworthy. For the question that asked the participants if they thought they were in an abusive relationship or not, 66.7% of the participants responded that they strongly disagreed although fourteen out of eighteen participants labeled the abusive text messages correctly. To understand this result further, we conducted a study to understand the effect of data visualization on user trust and what are the factors that people take into account when concluding they are in an abusive relationship. The next chapter discusses the method and results of the study.

Chapter 9

Understanding user perception and trust when results from a dating abuse detection application are displayed

9.1 Introduction

Chapter 5 evaluates the performance of various machine learning algorithms, and the success of the detection application of the phone app is heavily dependent on the accuracy of the classifier we are using. The performance of the Linear Support Vector Machine with a tf-idf feature extractor with a unigram input compared to the other classifiers is approximately 91%, which is in the same range of the other mental health-related apps reported in the literature. Due to the sensate nature of the information that is being presented to the user, it is also essential to be extremely cautious about how the results are displayed. We are interested in understanding how data visualization has an impact on a user's trust in the information reported by the app. The reasons behind exploring this paradigm are 1) how can we support the credibility of the detection algorithm to enable the user to trust the results they view and 2) is there a correlation between users trust in the results presented and their likelihood in seeking help. As the purpose of the phone app is to stop abuse preemptively in its early stages, user trust plays a huge role in the intervention and success of this app. Interpersonal relationships, what is acceptable in a relationship, and language usage are subjective and unique to every couple. Our training set is a representative sample but does not account for the diversity in language and relationship dynamics. We propose designing an user-interface that will empower the user to choose whether they want help or not. The purpose of the detection feature and the machine learning classifier is to support the decision-making process of whether the user is in an abusive relationship or not. We also want to explore the social question What is the threshold value of a user considering themselves in an abusive relationship, i.e., what are the number of text messages that would enable the user to consider themselves in an abusive relationship.

We conducted a user study to answer these two research questions

1. What is the threshold of abusive text messages that would motivate the user to consider themselves in an abusive relationship?

2. What is the most effective way to visualize the results of the detection classifier that would invoke user trust and encourage them to receive necessary help?

The following chapter discusses the study design and preliminary data analysis trends that would help us answer the research mentioned above questions.

9.2 Methods

9.2.1 Participants and study design

To gain an understanding related to our research questions described above, we conducted a cross-sectional survey. The survey was a within-subject study where we recruited 202 participants who were randomly assigned to one of the nine conditions (Table). In this section, we describe how we developed the survey, recruited participants, overview of the survey questionnaires, and study procedure Participant selection and Demographic Questionnaire As the target population of the phone app is college-aged women in the United States, we primarily recruited female participants who had a minimum U.S. High School Education as a representative sample to participate in this study. They were given a demographic questionnaire (Appendix D) and a modified dating abuse awareness questionnaire asking about their knowledge or personal experience related to interpersonal violence.

9.2.2 Survey Development, Conditions and Participant Recruitment

There were two primary research questions and three conditions to support to each of these questions -

1. Presented with a set of text messages which contained both abusive and non-abusive text messages

   (a) What is the threshold of abusive text messages (in a percentage value) that will lead a user to conclude they are in an abusive relationship? Three conditions are

      i. 50% -50% (abusive-non-abusive)

      ii. 70%-30% (abusive – non-abusive)

      iii. 30%-70% (abusive – non-abusive)

   (b) Does data visualization have an impact on the user's trust when it comes to using the detection application? Three conditions

      i. Pie chart showing a percentage value for abusive and non-abusive text messages.

      ii. Text messages that have been labeled as abusive and non-abusive.

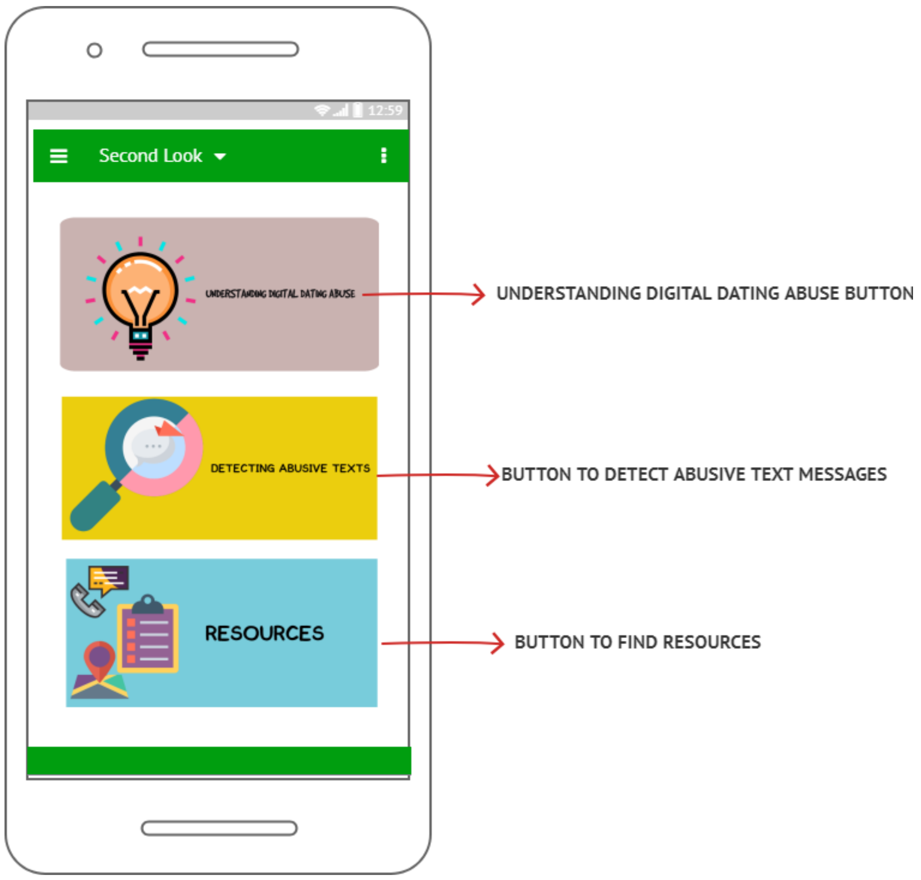
      iii. Pie chart showing the percentage value text messages that have been labeled as abusive and non-abusive.

The survey was developed using Balsamiq's online wireframe tool to create the user-interface mockups and deployed through Qualtrics.com. The participants were recruited through Mechanical Turk is a crowdsourcing web service that manages the supply and demand of tasks requiring human intelligence to complete. This service allows requesters to post small tasks called Human Intelligence Tasks (HITS). Workers selected HITS and completed in exchange for a small payment based on the requesters' set rate [42] [64]. We recruited 202 participants in all with approximately 21 (Mean=22.4, SD= 1.13) participants per condition due to the elimination of incomplete surveys.

Percentage of Abusive vs. non-Abusive Text messages	**Visualization 1 – Pie Chart only**	**Visualization 2 – Text Messages with corresponding labels only**	**Visualization 3 - Pie & Text**
50%-50%	Condition 1	Condition 4	Condition 7
70%-30%	Condition 2	Condition 5	Condition 8
30%-70%	Condition 3	Condition 6	Condition 9

9.2.3 Survey Questions

For all participants Questions 1 2 which described the experiment, introduced the home screen and provided them the definition of digital dating abuse was constant. (Figure 9.1, 9.2 and 9.3). The following questions were randomly selected from the nine different conditions to show the user-interface mockup of the detection page with the analyzed results. Figure 9.4a – 9i below illustrates all the nine conditions. The participants were then asked to answer the following questions using a 5-point Likert scale (Strongly Disagree – Strongly Agree) 1) If you were Jane, based on the results presented in Figure 4, would you consider yourself in an abusive dating relationship? 2) Have the results been presented in an effective manner such that you understand the analysis of the text messages clearly? 3a) If you were Jane, would you trust the information about the abusive text messages that the app has provided to you? 3b) Based on the definition of digital dating abuse provided earlier, please drag and drop the texts in the three boxes provided and arrange the text messages into three two categories (Abusive, Non-Abusive Not-Sure). All the participants were shown in Figure 9.5 which is the mock-up of the resources page and asked the following questions 1) In your opinion, would Jane use the resources feature as shown above to get help? 2) As a user, would you use the resources feature as shown above to get help? 3) Moreover, Rank the list of resources in order of preference (1- most preferred)

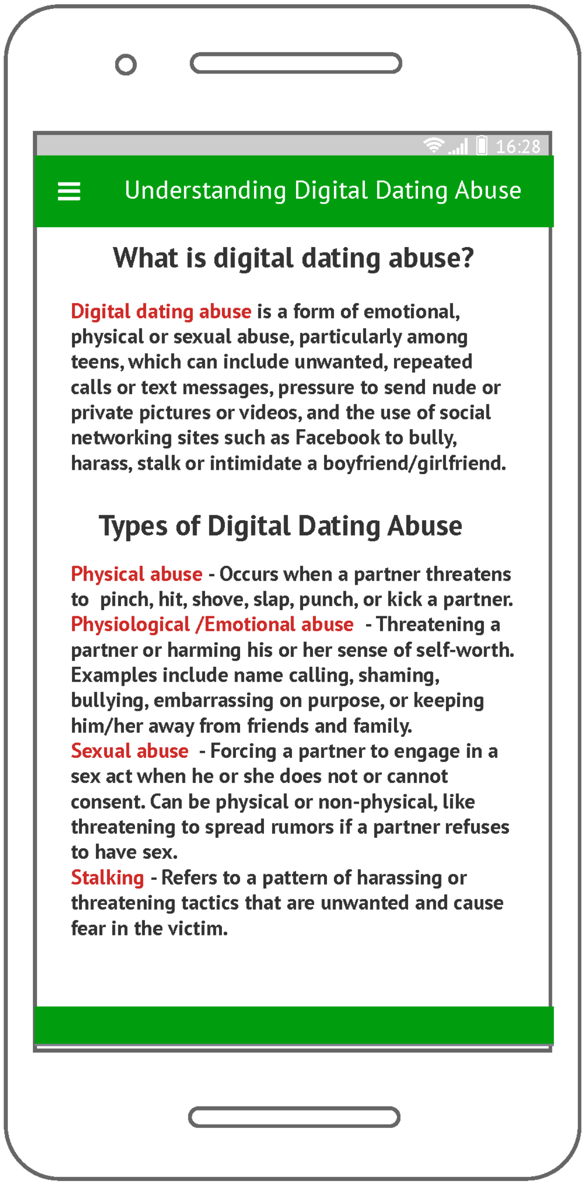


UNDERSTANDING DIGITAL DATING ABUSE BUTTON

BUTTON TO DETECT ABUSIVE TEXT MESSAGES

BUTTON TO FIND RESOURCES

Figure 9.1: illustrates the home screen or landing page of the app



Understanding Digital Dating Abuse

What is digital dating abuse?

Digital dating abuse is a form of emotional, physical or sexual abuse, particularly among teens, which can include unwanted, repeated calls or text messages, pressure to send nude or private pictures or videos, and the use of social networking sites such as Facebook to bully, harass, stalk or intimidate a boyfriend/girlfriend.

Types of Digital Dating Abuse

Physical abuse - Occurs when a partner threatens to pinch, hit, shove, slap, punch, or kick a partner.
Physiological /Emotional abuse - Threatening a partner or harming his or her sense of self-worth. Examples include name calling, shaming, bullying, embarrassing on purpose, or keeping him/her away from friends and family.
Sexual abuse - Forcing a partner to engage in a sex act when he or she does not or cannot consent. Can be physical or non-physical, like threatening to spread rumors if a partner refuses to have sex.
Stalking - Refers to a pattern of harassing or threatening tactics that are unwanted and cause fear in the victim.

Figure 9.2: illustrates the page containing the definition of digital dating abuse

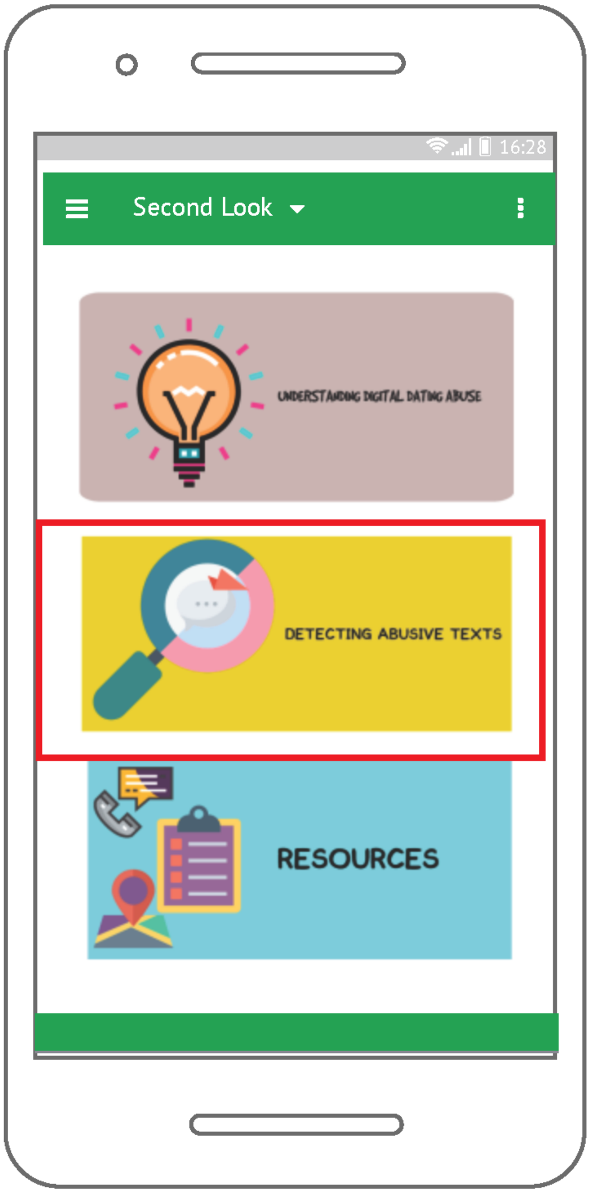
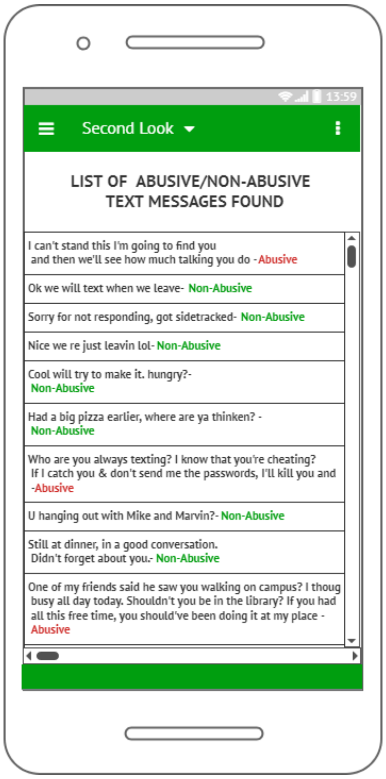


Figure 9.3: In a hypothetical scenario imagine you are Jane. To evaluate the text messages, Jack has sent you, click on the 'Detecting Abusive Text Messages Button' (To simulate this click, please click on the arrow at the bottom of the page



LIST OF ABUSIVE/NON-ABUSIVE TEXT MESSAGES FOUND

I can't stand this I'm going to find you and then we'll see how much talking you do -Abusive

Ok we will text when we leave- Non-Abusive

Sorry for not responding, got sidetracked- Non-Abusive

Nice we re just leavin lol- Non-Abusive

Cool will try to make it. hungry?- Non-Abusive

Had a big pizza earlier, where are ya thinken? - Non-Abusive

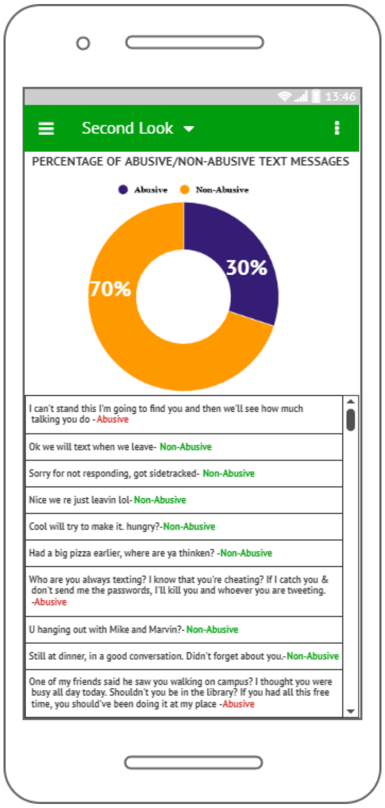
Who are you always texting? I know that you're cheating? If I catch you & don't send me the passwords, I'll kill you and -Abusive

U hanging out with Mike and Marvin?- Non-Abusive

Still at dinner, in a good conversation. Didn't forget about you.- Non-Abusive

One of my friends said he saw you walking on campus? I thoug busy all day today. Shouldn't you be in the library? If you had all this free time, you should've been doing it at my place - Abusive

Figure 9.4: illustrates the home screen or landing page of the app



PERCENTAGE OF ABUSIVE/NON-ABUSIVE TEXT MESSAGES

I can't stand this I'm going to find you and then we'll see how much talking you do -Abusive

Ok we will text when we leave- Non-Abusive

Sorry for not responding, got sidetracked- Non-Abusive

Nice we re just leavin lol- Non-Abusive

Cool will try to make it. hungry?-Non-Abusive

Had a big pizza earlier, where are ya thinken? -Non-Abusive

Who are you always texting? I know that you're cheating? If I catch you & don't send me the passwords, I'll kill you and whoever you are tweeting. -Abusive

U hanging out with Mike and Marvin?- Non-Abusive

Still at dinner, in a good conversation. Didn't forget about you.-Non-Abusive

One of my friends said he saw you walking on campus? I thought you were busy all day today. Shouldn't you be in the library? If you had all this free time, you should've been doing it at my place -Abusive

Figure 9.5: illustrates the home screen or landing page of the app

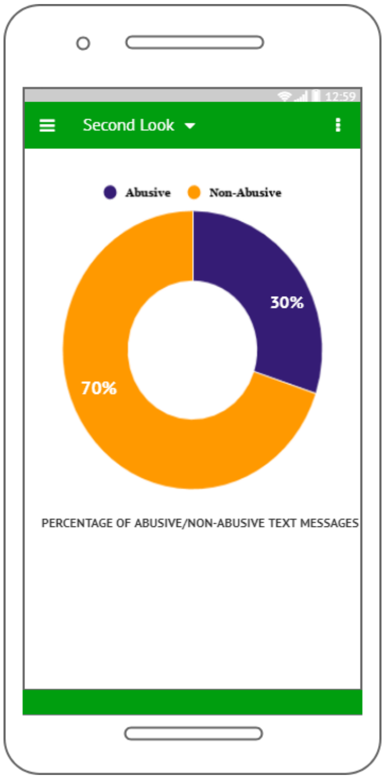


Figure 9.6: illustrates the home screen or landing page of the app

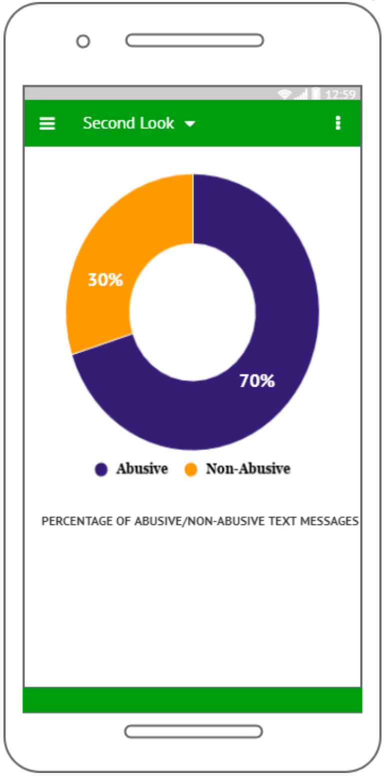
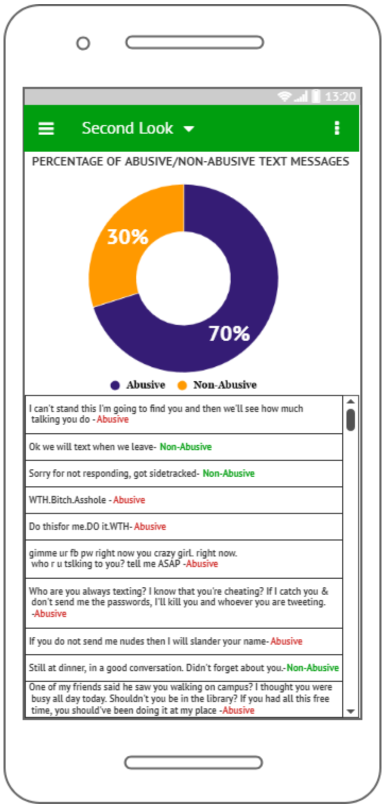


Figure 9.7: illustrates the home screen or landing page of the app



I can't stand this I'm going to find you and then we'll see how much talking you do -Abusive

Ok we will text when we leave- Non-Abusive

Sorry for not responding, got sidetracked- Non-Abusive

WTH.Bitch.Asshole -Abusive

Do thisfor me.DO it.WTH- Abusive

gimme ur fb pw right now you crazy girl. right now. who r u tslking to you? tell me ASAP -Abusive

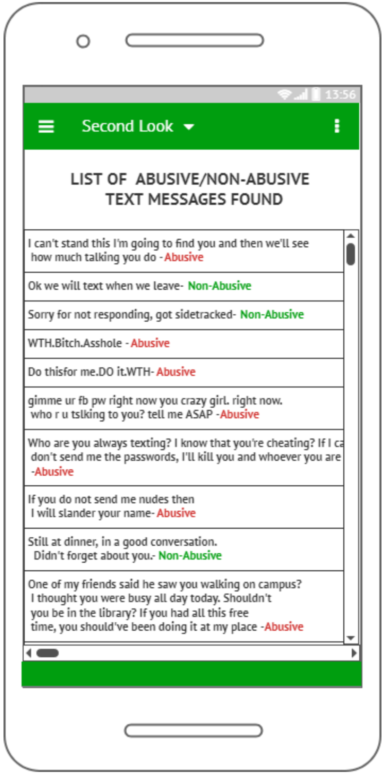
Who are you always texting? I know that you're cheating? If I catch you & don't send me the passwords, I'll kill you and whoever you are tweeting. -Abusive

If you do not send me nudes then I will slander your name- Abusive

Still at dinner, in a good conversation. Didn't forget about you.-Non-Abusive

One of my friends said he saw you walking on campus? I thought you were busy all day today. Shouldn't you be in the library? If you had all this free time, you should've been doing it at my place -Abusive

Figure 9.8: illustrates the home screen or landing page of the app



I can't stand this I'm going to find you and then we'll see how much talking you do -Abusive

Ok we will text when we leave- Non-Abusive

Sorry for not responding, got sidetracked- Non-Abusive

WTH.Bitch.Asshole -Abusive

Do thisfor me.DO it.WTH- Abusive

gimme ur fb pw right now you crazy girl. right now. who r u tslking to you? tell me ASAP -Abusive

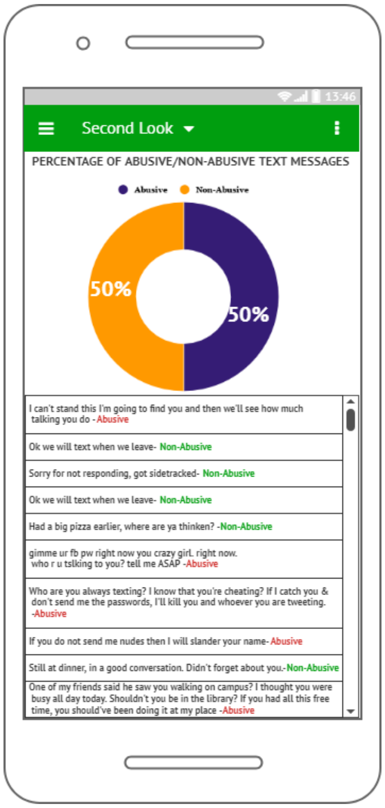
Who are you always texting? I know that you're cheating? If I ca don't send me the passwords, I'll kill you and whoever you are -Abusive

If you do not send me nudes then I will slander your name- Abusive

Still at dinner, in a good conversation. Didn't forget about you.- Non-Abusive

One of my friends said he saw you walking on campus? I thought you were busy all day today. Shouldn't you be in the library? If you had all this free time, you should've been doing it at my place -Abusive

Figure 9.9: illustrates the home screen or landing page of the app



PERCENTAGE OF ABUSIVE/NON-ABUSIVE TEXT MESSAGES

I can't stand this I'm going to find you and then we'll see how much talking you do - Abusive

Ok we will text when we leave- Non-Abusive

Sorry for not responding, got sidetracked- Non-Abusive

Ok we will text when we leave- Non-Abusive

Had a big pizza earlier, where are ya thinken? -Non-Abusive

gimme ur fb pw right now you crazy girl. right now. who r u tslking to you? tell me ASAP -Abusive

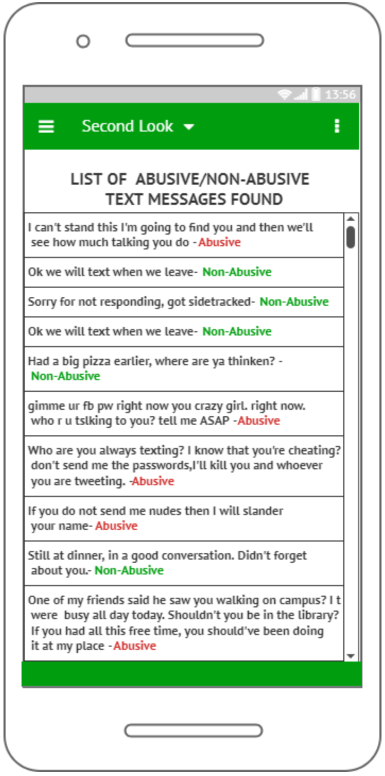
Who are you always texting? I know that you're cheating? If I catch you & don't send me the passwords, I'll kill you and whoever you are tweeting. -Abusive

If you do not send me nudes then I will slander your name- Abusive

Still at dinner, in a good conversation. Didn't forget about you.-Non-Abusive

One of my friends said he saw you walking on campus? I thought you were busy all day today. Shouldn't you be in the library? If you had all this free time, you should've been doing it at my place -Abusive

Figure 9.10: illustrates the home screen or landing page of the app



LIST OF ABUSIVE/NON-ABUSIVE TEXT MESSAGES FOUND

I can't stand this I'm going to find you and then we'll see how much talking you do - Abusive

Ok we will text when we leave- Non-Abusive

Sorry for not responding, got sidetracked- Non-Abusive

Ok we will text when we leave- Non-Abusive

Had a big pizza earlier, where are ya thinken? - Non-Abusive

gimme ur fb pw right now you crazy girl. right now. who r u tslking to you? tell me ASAP -Abusive

Who are you always texting? I know that you're cheating? don't send me the passwords,I'll kill you and whoever you are tweeting. -Abusive

If you do not send me nudes then I will slander your name- Abusive

Still at dinner, in a good conversation. Didn't forget about you.- Non-Abusive

One of my friends said he saw you walking on campus? I t were busy all day today. Shouldn't you be in the library? If you had all this free time, you should've been doing it at my place -Abusive

Figure 9.11: illustrates the home screen or landing page of the app

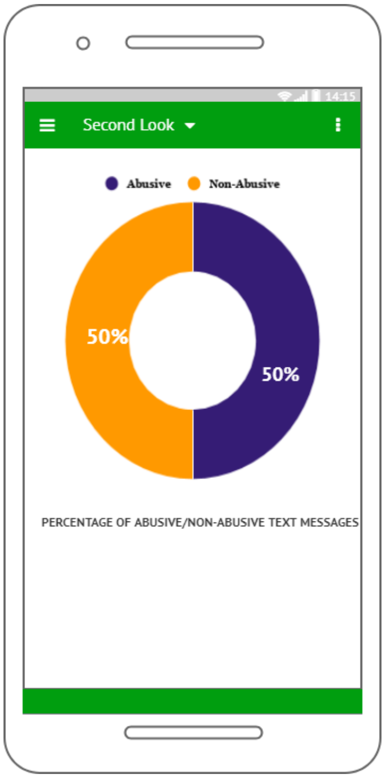
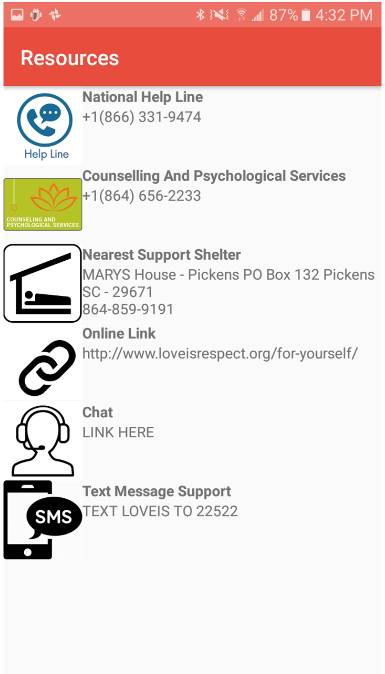


Figure 9.12: illustrates the home screen or landing page of the app



Resources

National Help Line
+1(866) 331-9474

Counselling And Psychological Services
+1(864) 656-2233

Nearest Support Shelter
MARYS House - Pickens PO Box 132 Pickens
SC - 29671
864-859-9191

Online Link
http://www.loveisrespect.org/for-yourself/

Chat
LINK HERE

Text Message Support
TEXT LOVEIS TO 22522

Figure 9.13: illustrates the home screen or landing page of the app

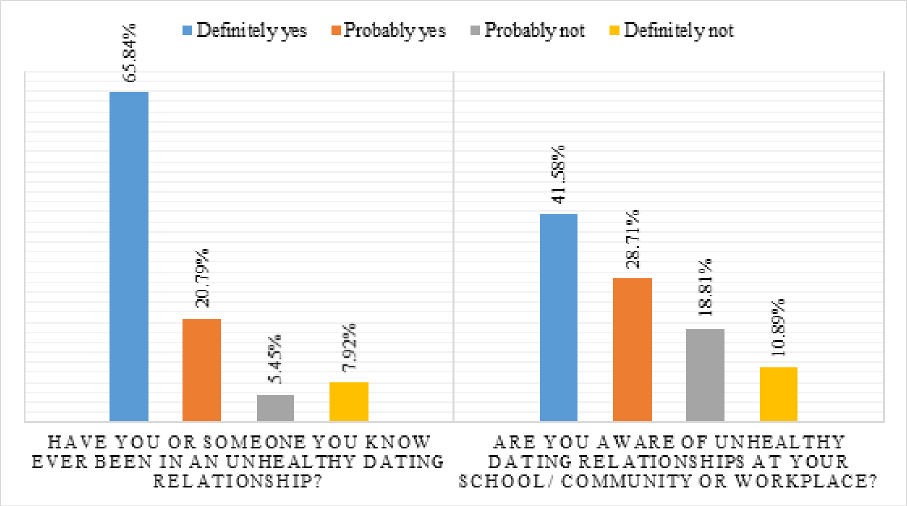


Figure 9.14: Illustrates the results of the demographic questionnaire

9.3 Results

9.4 Demographic Questionnaire

Table 9.2 illustrates the demographic information of all our participants. The mean age of all the participants was 36 years (SD=11.727), they were primarily female (96.53%), Caucasian (57.43%), had a bachelor's degree (42.57%) and married (48.51%). The participant pool is a representative of our target population regarding ethnicity, gender and education level. 65.5% of the participants selected Definitely yes when asked if they or someone they knew had ever been in an unhealthy relationship and 41.5% responded Definitely yes when asked if they were aware of unhealthy dating relationships at school/community or workplace. Both these responses reinforce the necessity of the phone app. In response to the question would you want to use technology, for example, phone or web application to help you understand if you are being abused, 50.50% replied yes, and 35.64% said maybe.

9.5 Data Analysis

From the survey, we collected Likert Scale responses from all the participants, which were coded as Strongly Agree, Somewhat Agree, Neither Agree or Disagree, Somewhat Disagree or Strongly Disagree. The responses were re-coded to numerical values 1 -5 where one was consid-

Gender	
Male	3.47%
Female	96.53%
Ethnicity	
Caucasian	57.43%
Hispanic	6.44%
African American or Black	10.89%
Native American or American Indian	2.48%
Asian / Pacific Islander	20.30%
Other	1.49%
Do not want to answer	0.99%
Education	
Some high school, no diploma	1.98%
High School graduate or the equivalent	11.88%
Some college credit no degree	22.77%
Sophomore	0.99%
Junior	1.98%
Freshman	0.00%
Senior	0.50%
Bachelor's Degree	42.57%
Master's Degree	17.33%
Marital Status	
Single	27.23%
Married	48.51%
Divorced	3.96%
In a relationship	16.83%
Others	2.97%
Do not wish to disclose	0.50%

Table 9.1: Demographic data for all participants

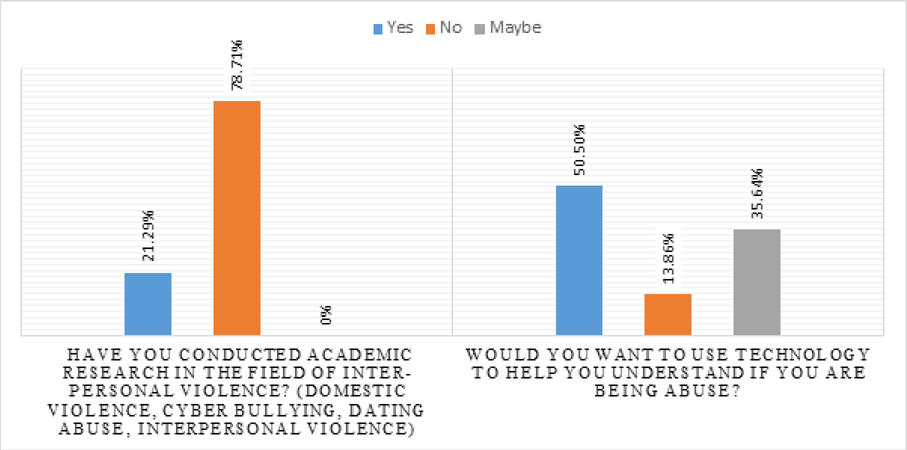


Figure 9.15: Illustrates the results of the demographic questionnaire

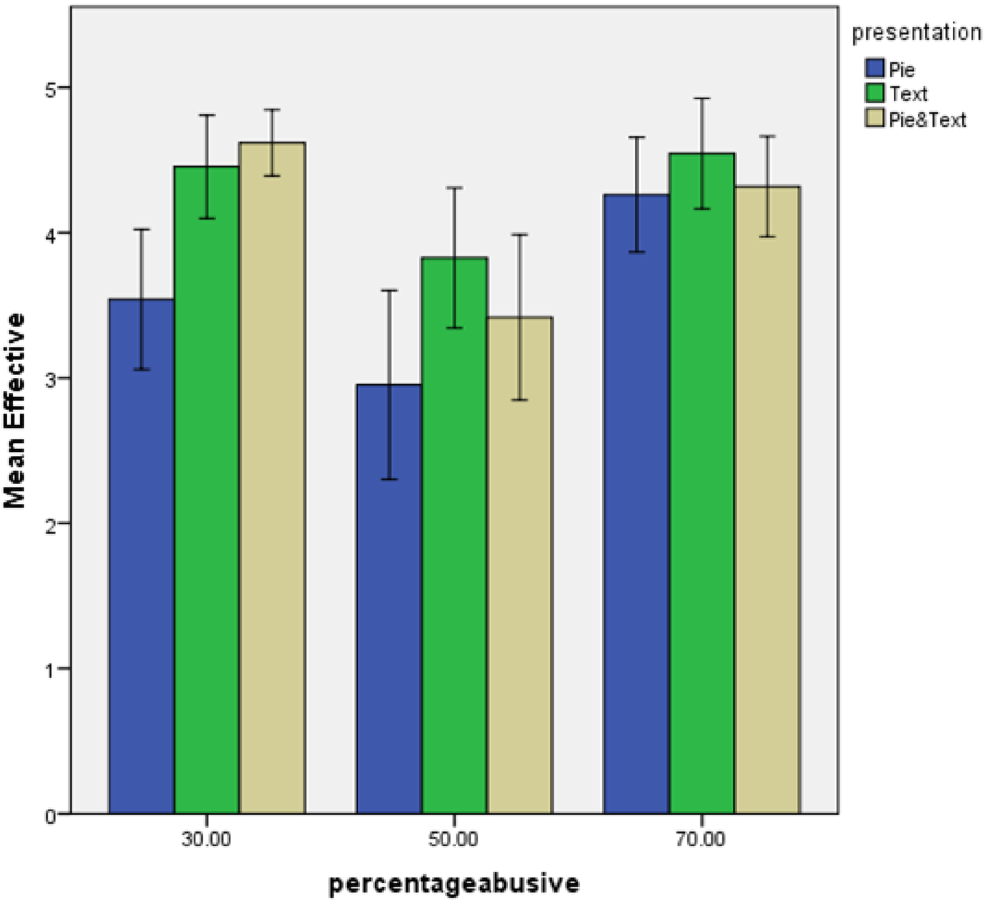
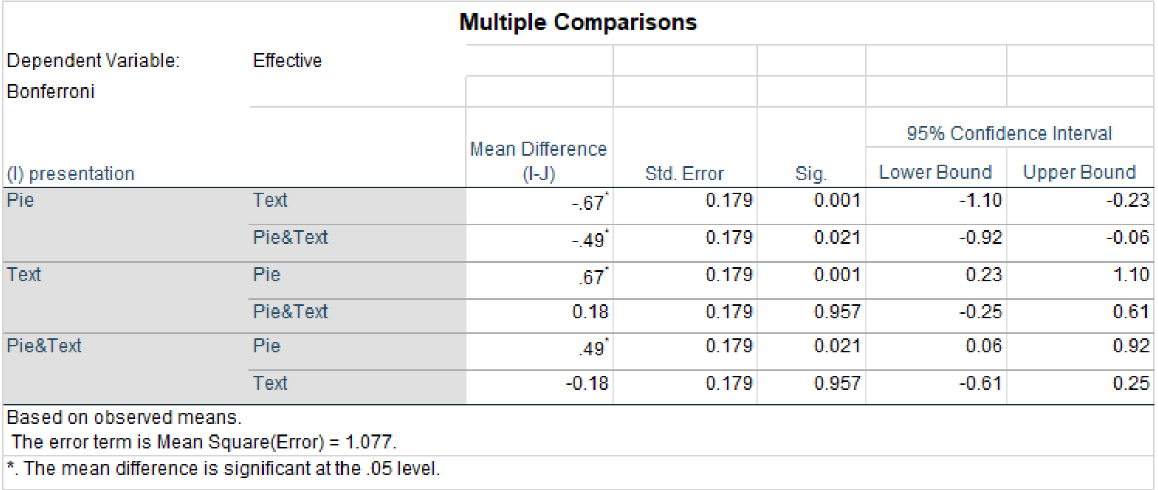


Figure 9.16: For the question *have the results been presented in an effective manner such that you understand the analysis of the text messages clearly*, comparing percentage abusive text messages, grouped by the three visualizations

ered as Strongly Agree and five was coded for Strongly Disagree. All further analysis was conducted on the re-coded numerical values.

We used SPSS to analyze each of the results for the questions mentioned in the previous section.We first conducted a full factorial ANOVA to determine if there is an interaction effect between any of the nine conditions (Visualizations - pie, text and pie & text and Percentage of abusive text messages - 30-70, 70-30 and 50 -50). As none of the factors had any interaction effect between each other, we did a uni variate posthoc multiple comparisons for observed means with Bonferonni correction[11].

For the questions *have the results been presented in an effective manner such that you understand the analysis of the text messages clearly?* we conducted multiple comparisons and across



Multiple Comparisons

Dependent Variable: Effective

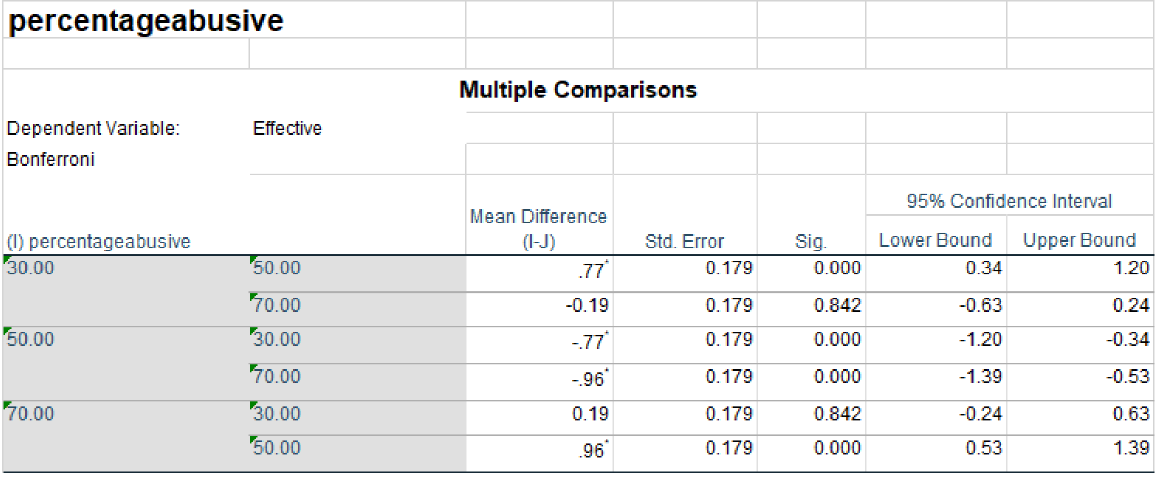
Bonferroni

(I) presentation		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Pie	Text	-.67*	0.179	0.001	-1.10	-0.23
	Pie&Text	-.49*	0.179	0.021	-0.92	-0.06
Text	Pie	.67*	0.179	0.001	0.23	1.10
	Pie&Text	0.18	0.179	0.957	-0.25	0.61
Pie&Text	Pie	.49*	0.179	0.021	0.06	0.92
	Text	-0.18	0.179	0.957	-0.61	0.25

Based on observed means.
The error term is Mean Square(Error) = 1.077.

*. The mean difference is significant at the .05 level.

Figure 9.17: For the question *have the results been presented in an effective manner such that you understand the analysis of the text messages clearly*, comparing visualizations



percentageabusive

Multiple Comparisons

Dependent Variable: Effective

Bonferroni

(I) percentageabusive		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
30.00	50.00	.77*	0.179	0.000	0.34	1.20
	70.00	-0.19	0.179	0.842	-0.63	0.24
50.00	30.00	-.77*	0.179	0.000	-1.20	-0.34
	70.00	-.96*	0.179	0.000	-1.39	-0.53
70.00	30.00	0.19	0.179	0.842	-0.24	0.63
	50.00	.96*	0.179	0.000	0.53	1.39

Figure 9.18: For the question *have the results been presented in an effective manner such that you understand the analysis of the text messages clearly*, comparing percentage abusive text messages

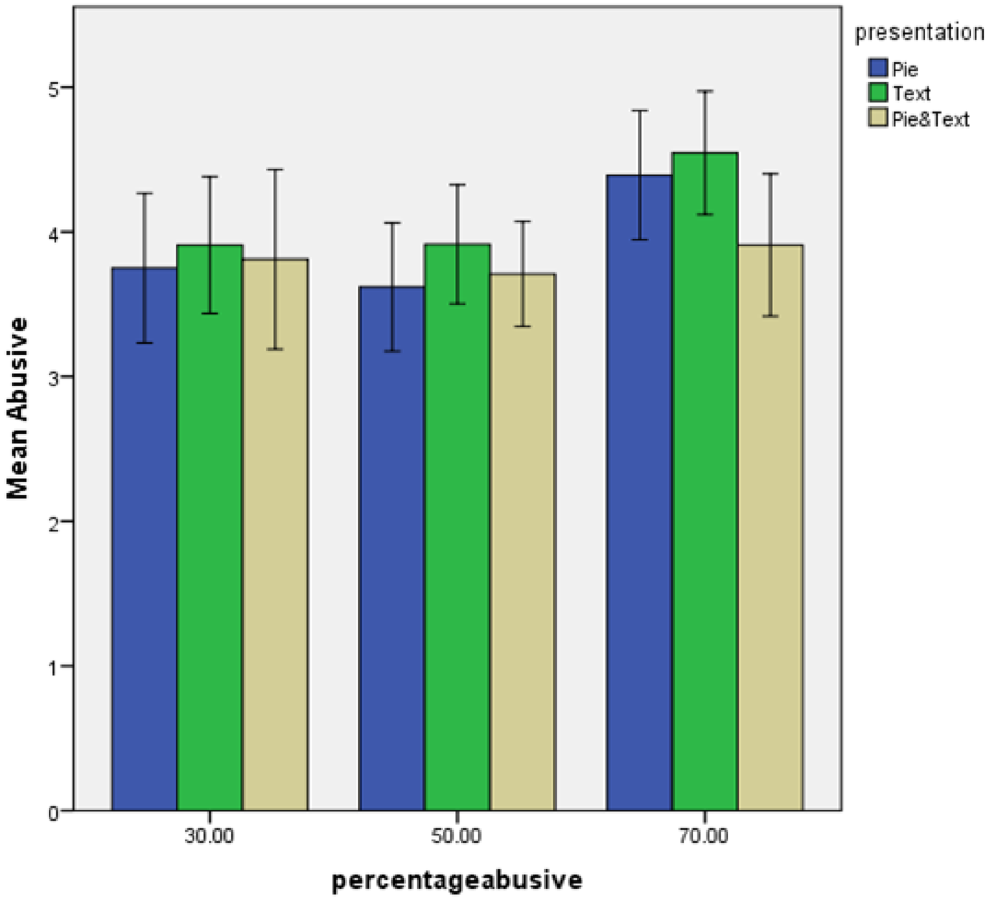
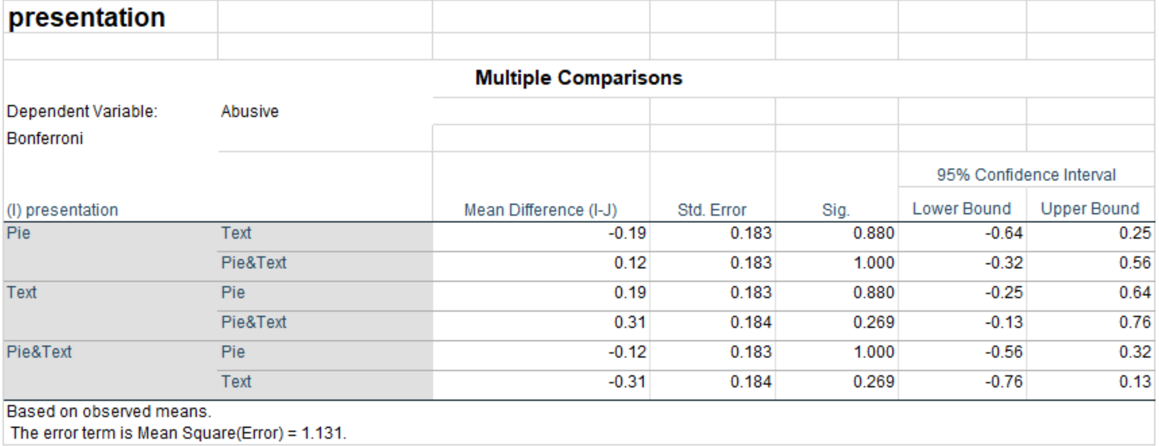


Figure 9.19: For the question *If you were Jane, based on the results, would you consider yourself in an abusive dating relationship?*, comparing comparing percentage abusive text messages, grouped by the three visualization conditions

three visualizations (text, pie and pie and text). We found significant mean difference between the conditions Pie vs Text (mean difference= -.67) and Pie vs Pie&Text (mean difference =-.49). There was no significance difference between the conditions Pie& Text and Text. Figure 9.17 illustrates the results for this question For the same question, we found significant difference between the conditions where the users where presented with the conditions 30% abusive vs 50% abusive (mean difference= 0.77) and conditions 70% abusive vs 50% abusive (mean difference$\bar{0}$.96). Figure 9.18 illustrates the results.

For the question ***If you were Jane, based on the results, would you consider yourself in an abusive dating relationship?*** , we did not find any significant difference when comparing the three different visualizations(Figure 9.20). When comparing percentage abusive vs



presentation

Multiple Comparisons

Dependent Variable: Abusive
Bonferroni

					95% Confidence Interval	
(I) presentation		Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
Pie	Text	-0.19	0.183	0.880	-0.64	0.25
	Pie&Text	0.12	0.183	1.000	-0.32	0.56
Text	Pie	0.19	0.183	0.880	-0.25	0.64
	Pie&Text	0.31	0.184	0.269	-0.13	0.76
Pie&Text	Pie	-0.12	0.183	1.000	-0.56	0.32
	Text	-0.31	0.184	0.269	-0.76	0.13

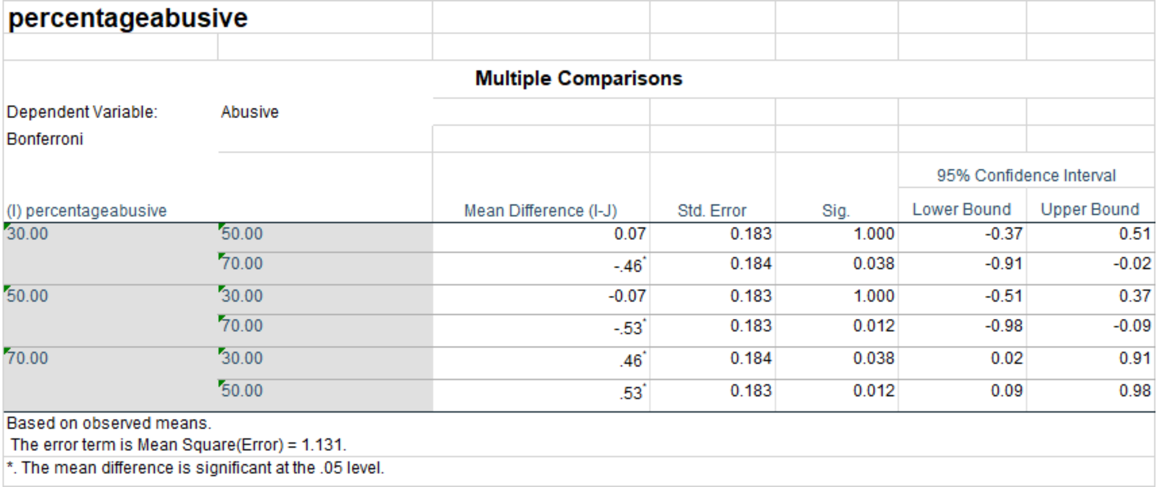
Based on observed means.
The error term is Mean Square(Error) = 1.131.

Figure 9.20: For the question *If you were Jane, based on the results, would you consider yourself in an abusive dating relationship?*, comparing visualizations

non-abusive text messages displayed to the user (Figure 9.21) we found significant differences between the conditions 70% vs 50% and 30% vs70%.

For the question ***If you were Jane, would you trust the information about the abusive text messages that the app has provided to you?*** , when comparing the three different visualizations(Figure 9.23). there was significant difference between the conditions Pie vs Text. When comparing percentage abusive vs non-abusive text messages displayed to the user (Figure 9.24) we found significant differences between the conditions 70% vs 50% and 30% vs50%.

For the questions **In your opinion, would Jane use the resources feature as shown above to get help?, comparing comparing percentage abusive text messages** and **As a user, would you use the resources feature as shown above to get help?** none of the conditions had any significant differences. Figures 9.26, 9.27 and 9.29, 9.30 illustrates the results.



percentageabusive

Multiple Comparisons

Dependent Variable: Abusive

Bonferroni

(I) percentageabusive		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
30.00	50.00	0.07	0.183	1.000	-0.37	0.51
	70.00	-.46*	0.184	0.038	-0.91	-0.02
50.00	30.00	-0.07	0.183	1.000	-0.51	0.37
	70.00	-.53*	0.183	0.012	-0.98	-0.09
70.00	30.00	.46*	0.184	0.038	0.02	0.91
	50.00	.53*	0.183	0.012	0.09	0.98

Based on observed means.
The error term is Mean Square(Error) = 1.131.

*. The mean difference is significant at the .05 level.

Figure 9.21: For the question *If you were Jane, based on the results, would you consider yourself in an abusive dating relationship?*, comparing comparing percentage abusive text messages

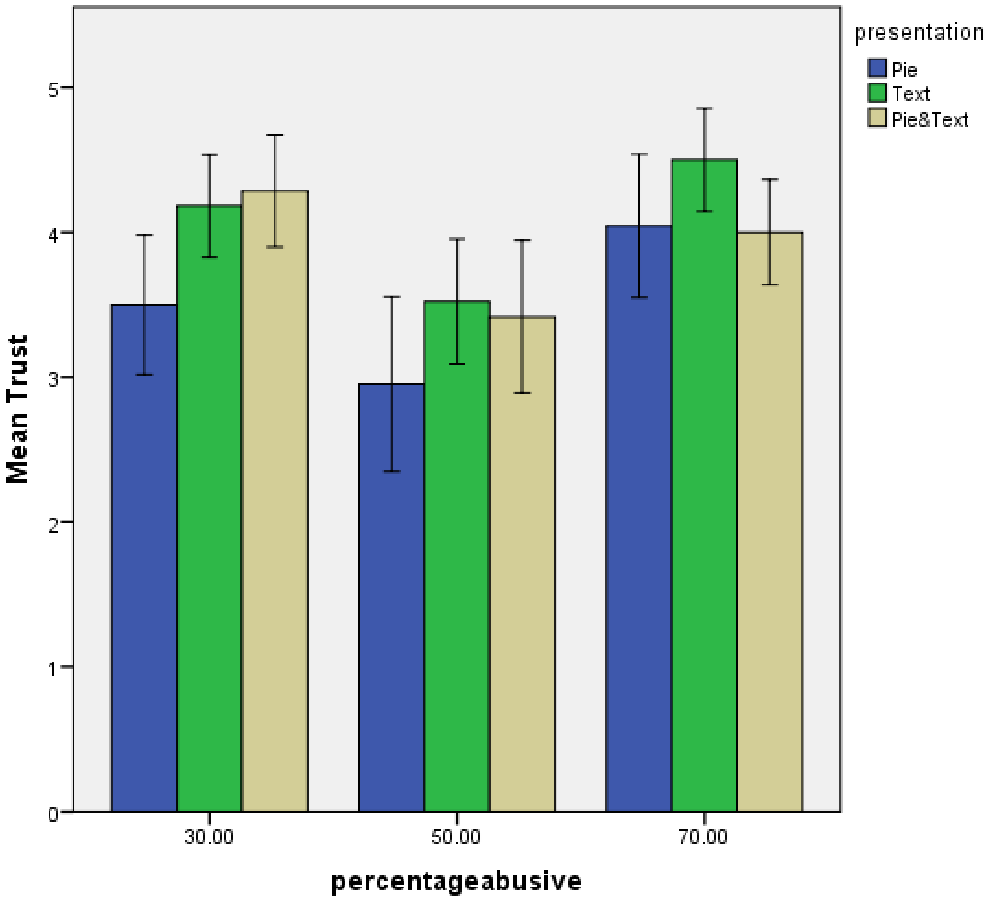
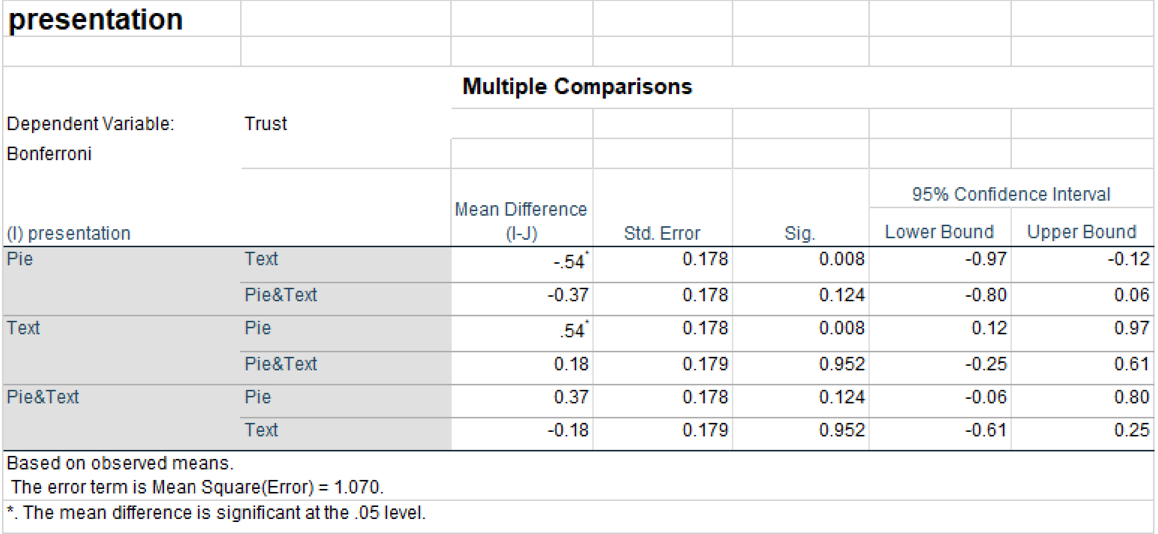


Figure 9.22: For the question *If you were Jane, would you trust the information about the abusive text messages that the app has provided to you?*, comparing percentage abusive text messages, grouped by the three visualization



presentation

Multiple Comparisons

Dependent Variable: Trust

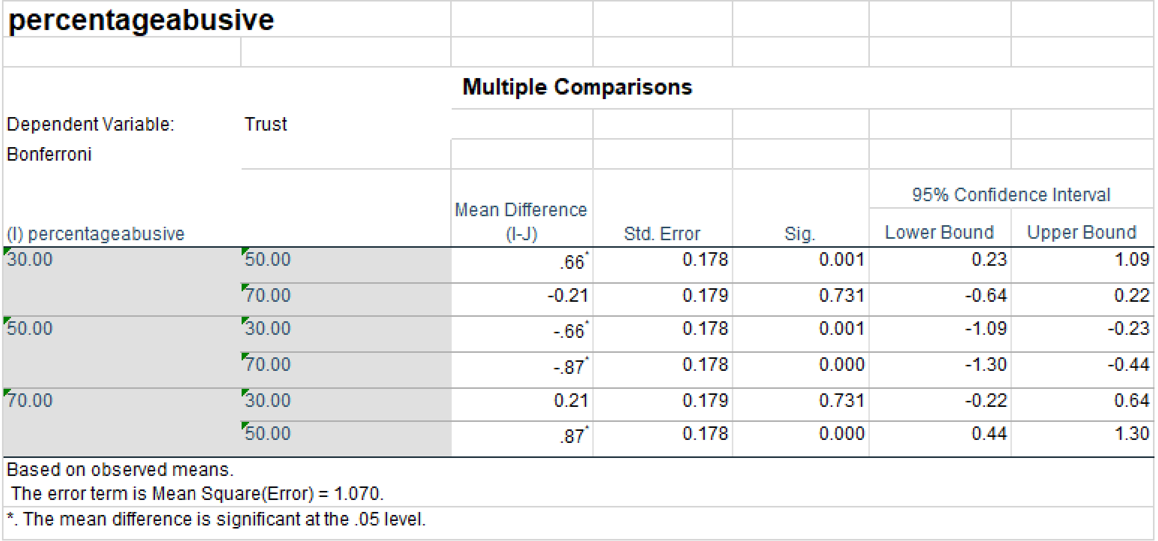
Bonferroni

(I) presentation		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Pie	Text	-.54*	0.178	0.008	-0.97	-0.12
	Pie&Text	-0.37	0.178	0.124	-0.80	0.06
Text	Pie	.54*	0.178	0.008	0.12	0.97
	Pie&Text	0.18	0.179	0.952	-0.25	0.61
Pie&Text	Pie	0.37	0.178	0.124	-0.06	0.80
	Text	-0.18	0.179	0.952	-0.61	0.25

Based on observed means.
The error term is Mean Square(Error) = 1.070.

*. The mean difference is significant at the .05 level.

Figure 9.23: For the question *If you were Jane, would you trust the information about the abusive text messages that the app has provided to you?*, comparing comparing percentage abusive text messages



percentageabusive

Multiple Comparisons

Dependent Variable: Trust

Bonferroni

(I) percentageabusive		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
30.00	50.00	.66*	0.178	0.001	0.23	1.09
	70.00	-0.21	0.179	0.731	-0.64	0.22
50.00	30.00	-.66*	0.178	0.001	-1.09	-0.23
	70.00	-.87*	0.178	0.000	-1.30	-0.44
70.00	30.00	0.21	0.179	0.731	-0.22	0.64
	50.00	.87*	0.178	0.000	0.44	1.30

Based on observed means.
The error term is Mean Square(Error) = 1.070.

*. The mean difference is significant at the .05 level.

Figure 9.24: For the question *If you were Jane, would you trust the information about the abusive text messages that the app has provided to you?*, comparing comparing percentage abusive text messages

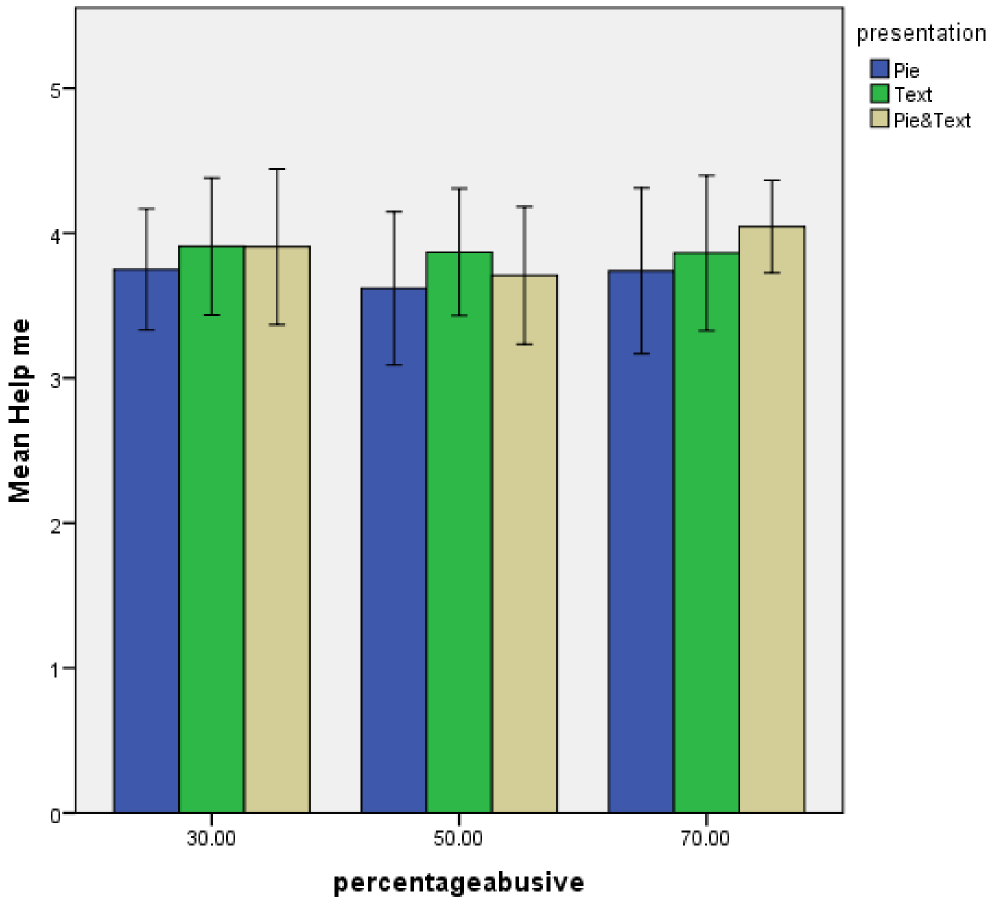
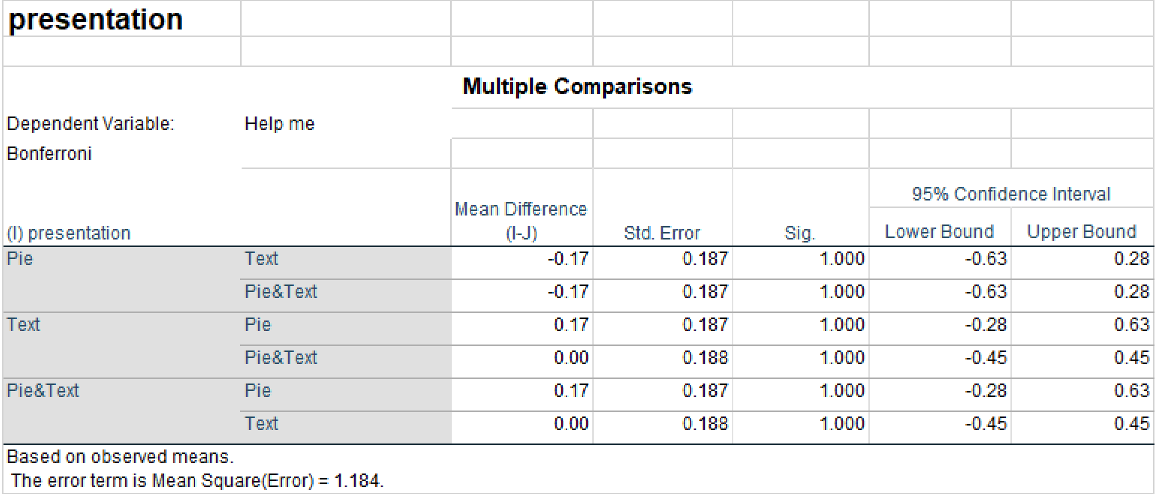


Figure 9.25: For the question *In your opinion, would Jane use the resources feature as shown above to get help?*, comparing percentage abusive text messages, grouped by the three visualizations



presentation

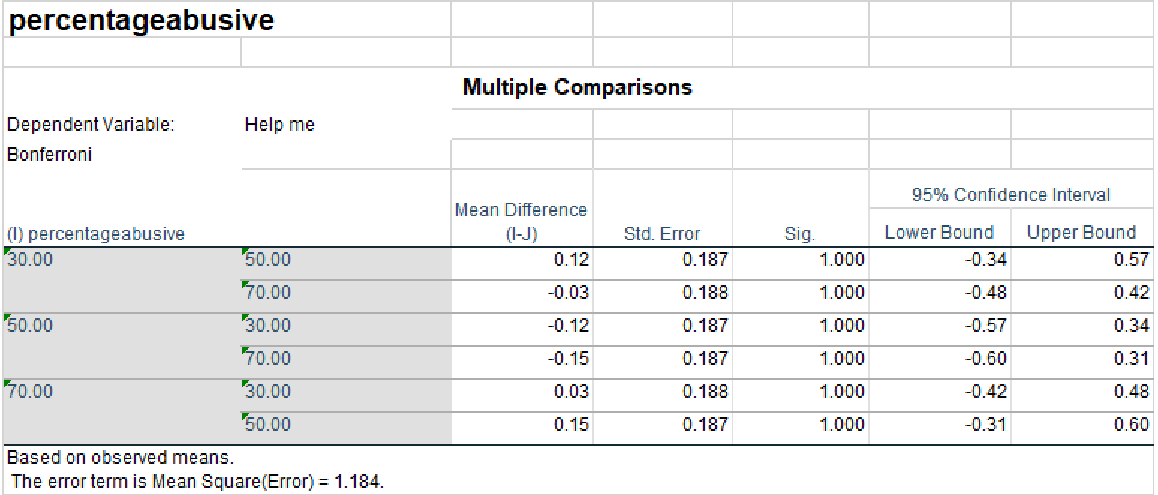
Multiple Comparisons

Dependent Variable: Help me
Bonferroni

(I) presentation		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Pie	Text	-0.17	0.187	1.000	-0.63	0.28
	Pie&Text	-0.17	0.187	1.000	-0.63	0.28
Text	Pie	0.17	0.187	1.000	-0.28	0.63
	Pie&Text	0.00	0.188	1.000	-0.45	0.45
Pie&Text	Pie	0.17	0.187	1.000	-0.28	0.63
	Text	0.00	0.188	1.000	-0.45	0.45

Based on observed means.
The error term is Mean Square(Error) = 1.184.

Figure 9.26: For the questions, In your opinion, would Jane use the resources feature as shown above to get help?,comparing visualizations



percentageabusive

Multiple Comparisons

Dependent Variable: Help me

Bonferroni

(I) percentageabusive		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
30.00	50.00	0.12	0.187	1.000	-0.34	0.57
	70.00	-0.03	0.188	1.000	-0.48	0.42
50.00	30.00	-0.12	0.187	1.000	-0.57	0.34
	70.00	-0.15	0.187	1.000	-0.60	0.31
70.00	30.00	0.03	0.188	1.000	-0.42	0.48
	50.00	0.15	0.187	1.000	-0.31	0.60

Based on observed means.
The error term is Mean Square(Error) = 1.184.

Figure 9.27: For the question, In your opinion, would Jane use the resources feature as shown above to get help?, comparing comparing percentage abusive text messages

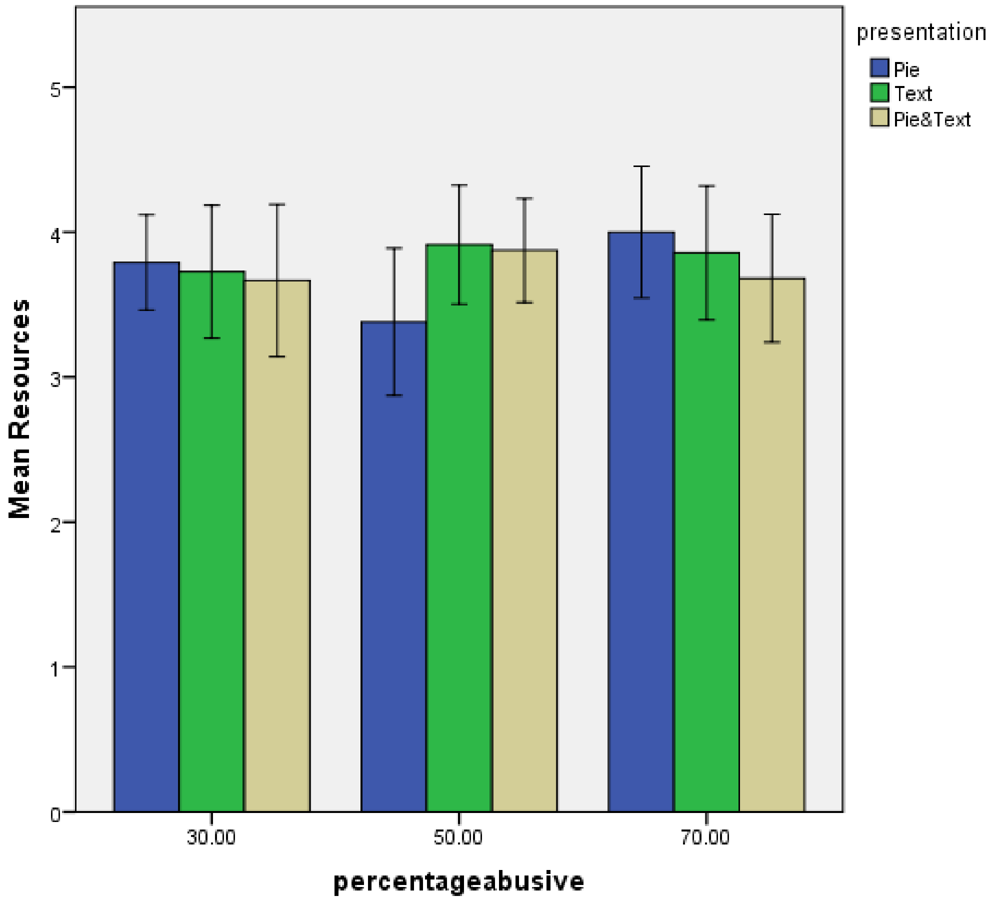
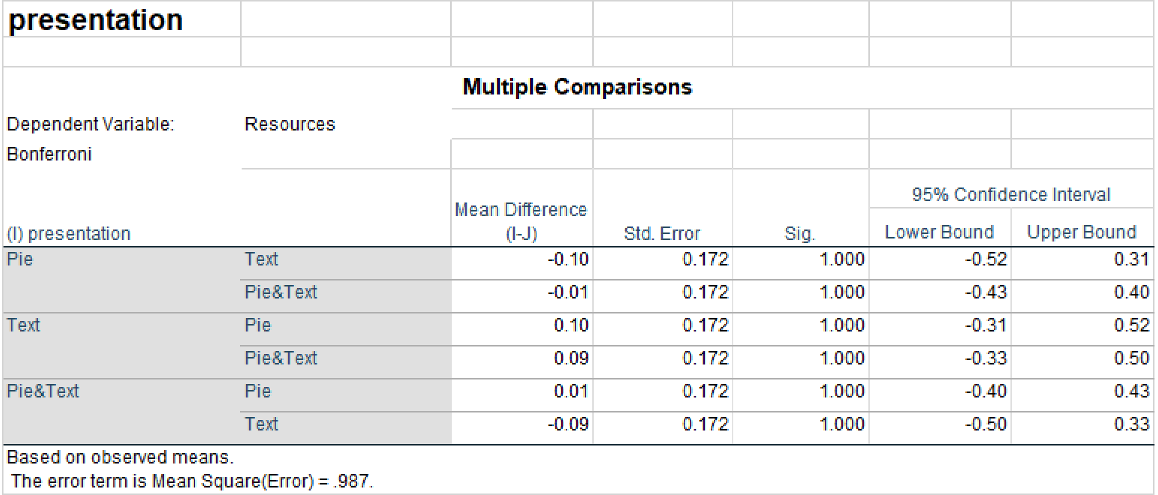


Figure 9.28: For the question *As a user, would you use the resources feature as shown above to get help?*, comparing percentage abusive text messages, grouped by the three visualizations



presentation

Multiple Comparisons

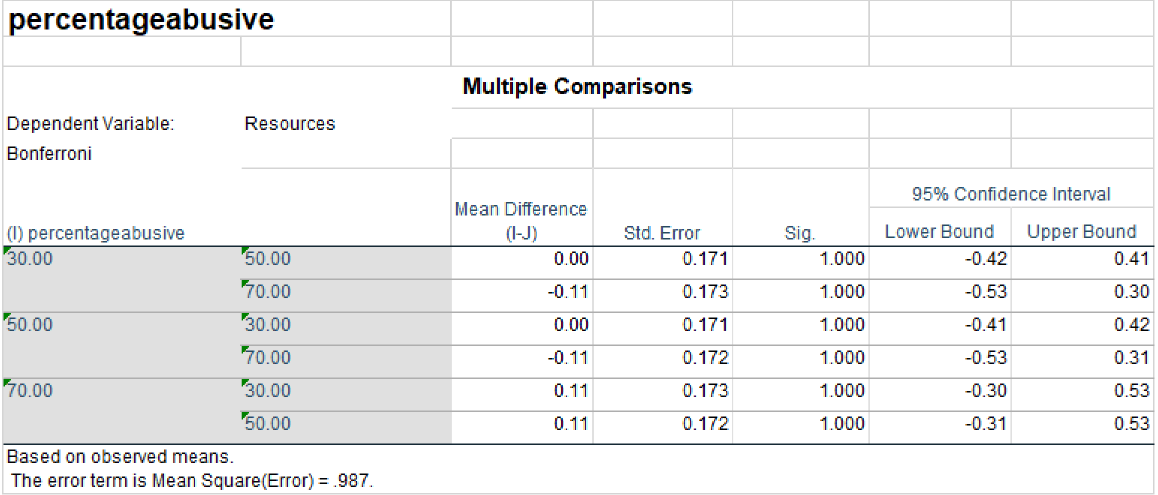
Dependent Variable: Resources

Bonferroni

(I) presentation		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Pie	Text	-0.10	0.172	1.000	-0.52	0.31
	Pie&Text	-0.01	0.172	1.000	-0.43	0.40
Text	Pie	0.10	0.172	1.000	-0.31	0.52
	Pie&Text	0.09	0.172	1.000	-0.33	0.50
Pie&Text	Pie	0.01	0.172	1.000	-0.40	0.43
	Text	-0.09	0.172	1.000	-0.50	0.33

Based on observed means.
The error term is Mean Square(Error) = .987.

Figure 9.29: Illustrates the results of the demographic questionnaire



percentageabusive

Multiple Comparisons

Dependent Variable: Resources

Bonferroni

(I) percentageabusive		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
30.00	50.00	0.00	0.171	1.000	-0.42	0.41
	70.00	-0.11	0.173	1.000	-0.53	0.30
50.00	30.00	0.00	0.171	1.000	-0.41	0.42
	70.00	-0.11	0.172	1.000	-0.53	0.31
70.00	30.00	0.11	0.173	1.000	-0.30	0.53
	50.00	0.11	0.172	1.000	-0.31	0.53

Based on observed means.
The error term is Mean Square(Error) = .987.

Figure 9.30: Illustrates the results of the demographic questionnaire

9.6 Discussion

The preliminary data trends indicate that when presented with a set of text messages with the percentage of abusive vs non-abusive text messages range 30%, 50% and 70% for the question *if you were Jane would you consider yourself in an abusive relationship*, there was significant difference between 30% abusive vs 70% and 50% vs 70% conditions respectively.For the thresholds mentioned above, we used three different visualizations a) Text messages- the labeled text messages displayed to the user b) Pie chart only - percentage abusive vs. non-abusive text messages displayed to the user c) Text messages and Pie chart - Displaying both the percentage of abusive vs. non-abusive text messages and labeled text messages.

Our hypothesis was the Text message the only condition would outperform the Pie and Text and Pie chart condition. From our analysis, we were unable to find a significant difference amongst any of these conditions.When considering data visualization and the user is trusting the information we found significant differences between the conditions Pie vs. Text which was aligned with our initial hypothesis. We hypothesized that if the user is presented with the labeled text messages, they will tend to trust the results rather than only depending on percentage output values presented from an unknown black box algorithm. We were unable to find any significant differences between the Text only vs. Pie and Text condition.

Based on this preliminary finding we propose extending this study and this is discussed in the Future work section (Chapter 10). The broader research impact of this study is further discussed in Chapter 12.

Chapter 10

Limitations and Future Work

This phone app is a proof of concept prototype promoting a mobile phone intervention to stop digital dating abuse. Throughout the process of development of this app, we have conducted user evaluation studies both on the usability of the app in general and the accuracy of the detection application. Listed below are certain limitations of the app and our future direction that we will be taking to addressing these limitations.

1. Design iteration of the android phone app We used a participatory design process in creating the android phone app, we received user and expert feedback on the design of the user-interface and we have updated the app based on those reviews. During the final usability evaluation where the participants interacted with the Android prototype we received feedback and suggestion on two specific issues which we will address in the next iteration of the android app.

   (a) Lack of navigation flexibility - Most of the participants pointed out the lack of a back button or navigation side bar which would allow them to access all the features without having to go to the landing page. We will be incorporating this side navigation bar into our app for the next iteration.

   (b) Analyzed text message is difficult to read- The participants also pointed out how that the layout of the analyzed text messages was not very readable. We propose to create a more readable and user friendly layout with color coded labels that would enable the user to understand this information at a quicker pace.

    (c) Remove the design redundancy on the landing page - We also received feedback regarding the design of the landing page of the app (the redundant buttons and clickable image), we would redesign this page to eliminate conflicting tasks.

    (d) Enable the links on the Resources page – For this current version the user's are unable to directly use the links on the resources page as it is a static display. We will be making these icons clickable to let the user's call the help line or access a web-page link directly from the app.

2. Increasing the size and linguistic diversity of the data-set for training machine learning algorithms Due to the sensitive nature of this problem space collecting abusive text messages is a challenge. Hence one of the limitations of this research is that the size of the initial data-set set consisting of both training and testing set is not adequate to capture all the features related to dating abuse. As a next step we plan to crowd source text messages using Mechanical Turk to increase the size and linguistic diversity of the training set. This would help us improve the validity of this method and establish a training set that can be used by future researchers.

3. For this dissertation, we only analyzed results related to binary labeling of text messages as abusive or non-abusive. For future work, we will be analyzing the classification of abusive sentences into different categories of abuse and use classifiers for multi-class classification. We will also explore other methods such as neural network and clustering techniques and compare the results with our current findings.

4. Exploring the visualization paradigm in supporting machine learning algorithms to increase user trust. Visualization paradigms are another research aspect that we will be exploring in the future in order to investigate different visualization techniques that can be used for conveying the information produced by the classifiers effectively to the user. The results from the preliminary study described in Chapter 9 is a step towards exploring this important paradigm especially in the field of mental health intervention apps. For future work we will be increasing our sample size for each condition (N=40) for a larger effect size and analyze the data to observe if the existing trends continue to be significant for a larger dataset.

5. Digital dating abuse is an important issue for young adults of all genders and sexual orientation. For this current project, we only explore females being abused in a heterosexual relationship.

With a more diverse dataset we want to extend, the reach of this application to other groups as well and help young adults make informed decisions about their relationships.

Chapter 11

Conclusion

Designing, building and evaluating SecondLook was motivated by the following research questions. Our goals were to -

1. Are the individual features of the app i.e. Awareness, Detection, and Resources designed in an intuitive manner that helps the user acquire the information they require?

2. Presented with a set of text messages which contains both abusive and non-abusive text messages

3. What is the threshold of abusive text messages (in percentage or count) to be considered in an abusive relationship?

4. Does the way the data is visualized have an impact on user's trust when it comes to using the detection application?

5. Is there a change in the user's perception of abusive vs. non-abusive in observing the text messages classified by the machine learning algorithm (as displayed on the app) vs. labeling the text messages themselves?

6. What is the threshold of abusive text messages that encourages users to seek help and if they do, which is their most preferred resources?

For question 1) We used a modified Heuristic Evaluation Approach and provided the users with specific tasks to evaluate the individual features of the phone app. We received feedback from 18 participants out of which majority were female participants. Overall usability of the phone application was rated to be between cosmetic usability problem to minor usability issues. The primary issues have been listed in Chapter 10.

When presented with a set of text messages with the percentage of abusive vs non-abusive text messages range 30%, 50% and 70% for the question *if you were Jane would you consider yourself in an abusive relationship*, there was significant difference between 30% abusive vs 70% and 50% vs 70% conditions respectively. We hypothesized that irrespective of the percentage of abusive text messages, the user would consider themselves in an abusive relationship, but from the preliminary data analysis, we conducted we were not able to obtain a clear answer to that question. We were also interested in understanding whether data visualization or amount of information provided to the user related to the results of the machine learning algorithm affected the user's perception of being in an abusive relationship.

For the thresholds mentioned above, we used three different visualizations a) Text messages- the labeled text messages displayed to the user b) Pie chart only - percentage abusive vs. non-abusive text messages displayed to the user c) Text messages and Pie chart - Displaying both the percentage of abusive vs. non-abusive text messages and labeled text messages. Our hypothesis was the Text message the only condition would outperform the Pie and Text and Pie chart condition. From our analysis, we were unable to find a significant difference amongst any of these conditions.

When considering data visualization and the user is trusting the information we found significant differences between the conditions Pie vs. Text which was aligned with our initial hypothesis. We hypothesized that if the user is presented with the labeled text messages, they will tend to trust the results rather than only depending on percentage output values presented from an unknown black box algorithm. We were unable to find any significant differences between the Text only vs. Pie and Text condition.

We were unable to find any significant difference amongst the three thresholds of abusive vs. non-abusive text messages that would motivate the user to get help. Amongst our 18 participants, the preferred resources were online resources such as websites, forums. We will use this information to restructure our resources page on the phone app to allow the users to get the most preferred resource on top.

We wanted to understand if there is a difference between the user self-labeling abusive text messages vs. rely on the machine learning results. We asked 18 of our participants to complete the task, we found that majority of the participants were able to label the messages correctly, but when they read the results did not conclude themselves to be in an abusive relationship. This result was of great significance to us, as it sheds light on the psychology of perception.

Although we were unable to draw concrete conclusions from all our research questions, we were able to see certain useful trends related to broader impact questions such as user trust in machine learning algorithms, the role of data visualizations in impacting user trust and societal and cultural perceptions as to what can be considered as an abusive relationship. These results also allude to the complexity of this research domain, and for future work, we propose to run more extensive and diverse participant groups for concrete answers to the research questions.

Chapter 12

Contributions and Broader Imapct

12.1 Contributions

1. Creating SecondLook : a prototype mobile intervention for digital dating abuse with three features

2. Created a detection application that uses the combination of a Linear Support Vector Machine, 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%.

3. Creating and validating a dataset of abusive text messages that can be used by future researchers as a training/testing dataset. A total of 161 abusive text messages are part of this dataset. We also created three additional balanced datasets with these abusive text messages and non-abusive text messages from two different publicly available datasets. We tested the robustness of our classifier on all three datasets, and the results were in the same range on all three datasets.

4. Designed, built and conducted a usability study of an android prototype app with the following features-

    (a) That allows users to understand more about healthy relationships regarding Characteristics of Healthy Relationships vs. unhealthy and abusive ones. This is provided through words that can be associated with these relationships and example scenarios. The awareness feature framework has been developed to be portable and easily expandable. As the information is hosted on a Google Drive, it's relatively easy to update or add new information material.

(b) We also provide users with an FAQ sheet (part of the same framework used for 1.3.1) which educates the user about global impact and deadly nature of dating abuse. It also serves the purpose of allowing the user who may be a victim of dating abuse to realize they are not alone in this and enable them to get help. itemThe app contains a mobile-friendly version of the Healthy Relationships Quiz which is available for the user on loveisrespect.org as a web page or a printable document. The app version allows the user to complete the quiz more discretely and the scoring rubric is programmed into this feature thus eliminating the user having to use the printable documents self-scoring rubric.

(c) The backend of the phone app is hosted on a python-based server called Flask, which is not computationally intensive compared to other servers. The backend framework (classifier and training set) is modular, so it can be updated/expanded without dismantling the working framework. This framework can also be used for other apps that use machine learning to understand patterns about a user's phone data.

12.2 Broader Impact

The broader impact of this research is not only limited to the immediate research domain of digital dating abuse. We created a framework and designed a modular android phone app that can be modified to analyze digital content on mobile phones to detect mental health issues such as user stress, depression, financial management issues and interpersonal violence-related issues such as workplace and racial harassment.

The framework that has been developed relies on three principles – awareness, detection and help resources. This framework can be expanded to address a broad domain of issues related to mental health, wellbeing and personal safety. It is important to know that oneself is in the middle of an issue, but it is equally important to understand the particular issue and have access to resources to combat the problem. Machine Learning has become an integral part of our everyday life and interactions with technology. They are being used in applications in health-care to detect cancer from imaging data to recommending movies and restaurants. There has always been a form of distrust between humans and technology and with the increase in the use of machine learning algorithms in sensitive areas such as healthcare and recommender systems we need to understand how to design systems that use machine learning algorithms and also motivate users to trust these technologies. This was extremely critical for our phone app, as the backbone of this app uses machine learning to address a sensitive and subjective topic such as dating abuse and relationships. This research explores data visualizations, the effect it has on user trust on the results presented, and we have seen promising trends with visualizations which show how the classifier worked vs. displaying only the final results from the machine learning algorithm.

Appendices

Appendix A

SPAM DATASET

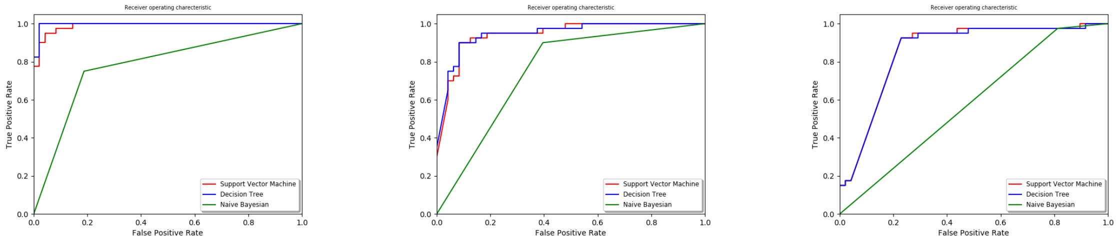


Figure A.1: ROC curve all three classifiers with unigram, bigram and trigram input respectively using tf-idf feature extractor

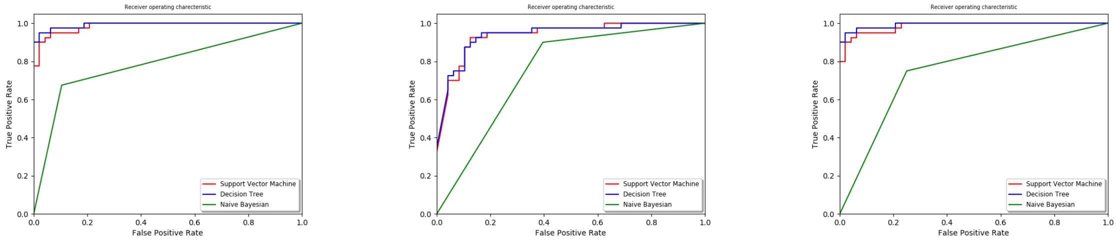


Figure A.2: ROC curve all three classifiers with the combination of unigram, bigram and trigram input respectively using tf-idf feature extractor

SPAM DATASET

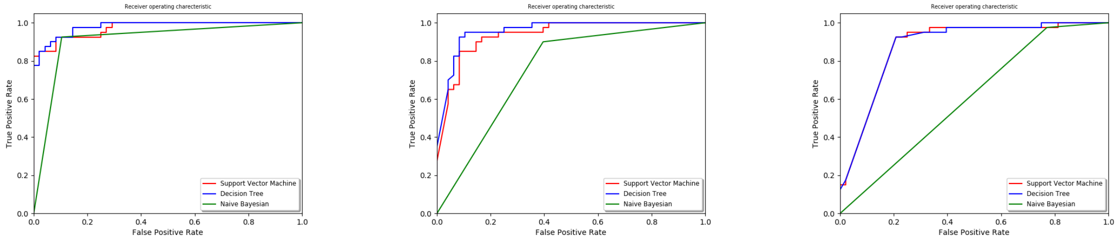


Figure A.3: ROC curve all three classifiers with unigram, bigram and trigram input respectively using count feature extractor

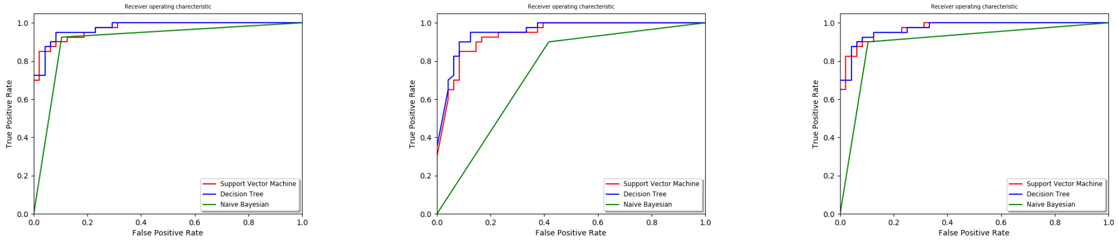


Figure A.4: ROC curve all three classifiers with the combination of unigram, bigram and trigram input respectively using count feature extractor

CRITICAL DATASET

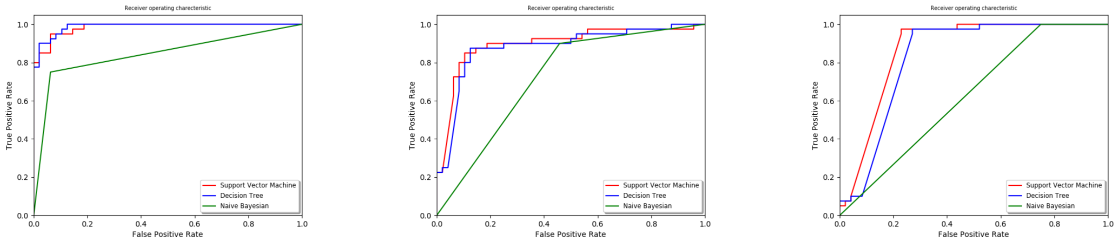


Figure A.5: ROC curve all three classifiers with unigram, bigram and trigram input respectively using tf-idf feature extractor

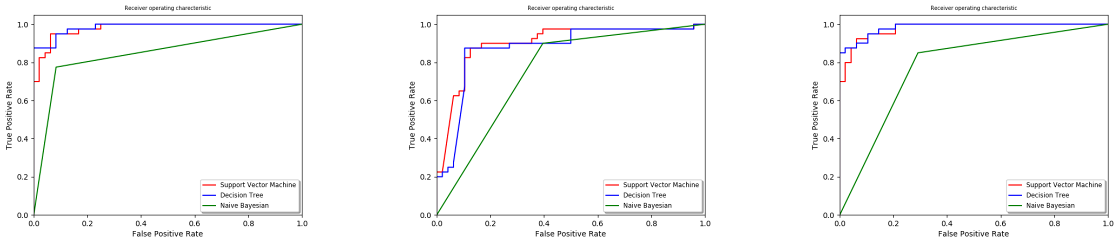


Figure A.6: ROC curve all three classifiers with the combination of unigram, bigram and trigram input respectively using tf-idf feature extractor

CRITICAL DATASET

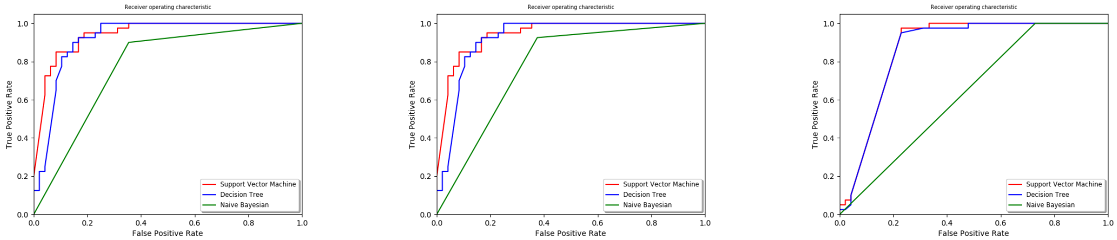


Figure A.7: ROC curve all three classifiers with unigram, bigram and trigram input respectively using count feature extractor

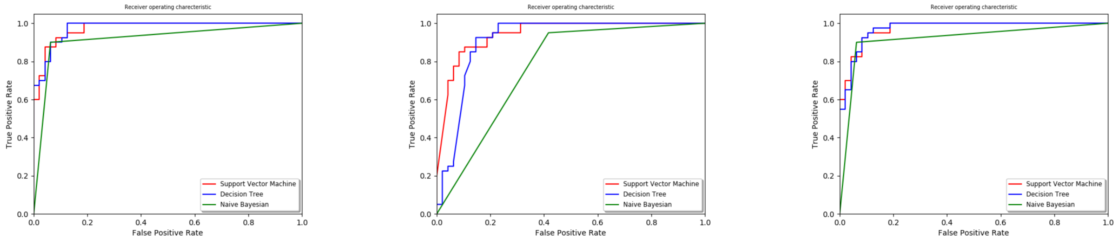


Figure A.8: ROC curve all three classifiers with the combination of unigram, bigram and trigram input respectively using count feature extractor

MIXED DATASET

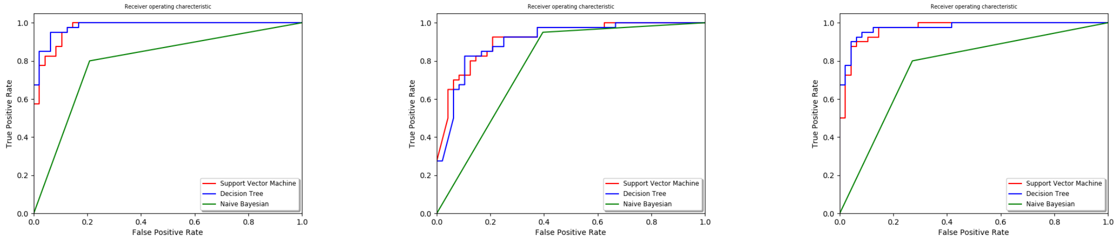


Figure A.9: ROC curve all three classifiers with unigram, bigram and trigram input respectively using tf-idf feature extractor

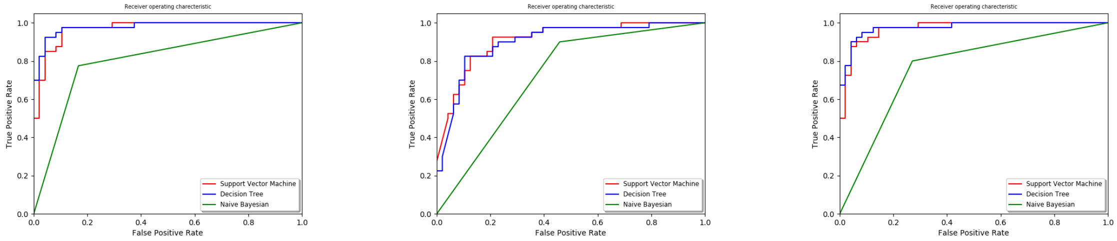


Figure A.10: ROC curve all three classifiers with the combination of unigram, bigram and trigram input respectively using tf-idf feature extractor

MIXED DATASET

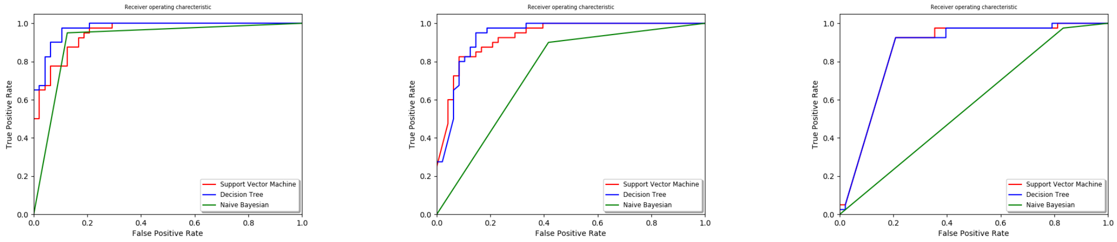


Figure A.11: ROC curve all three classifiers with unigram, bigram and trigram input respectively using count feature extractor

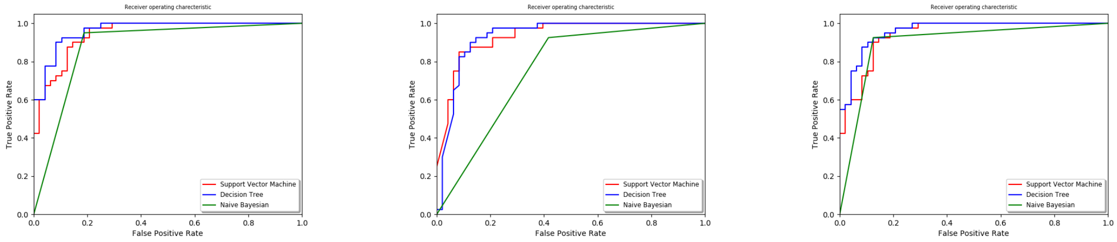
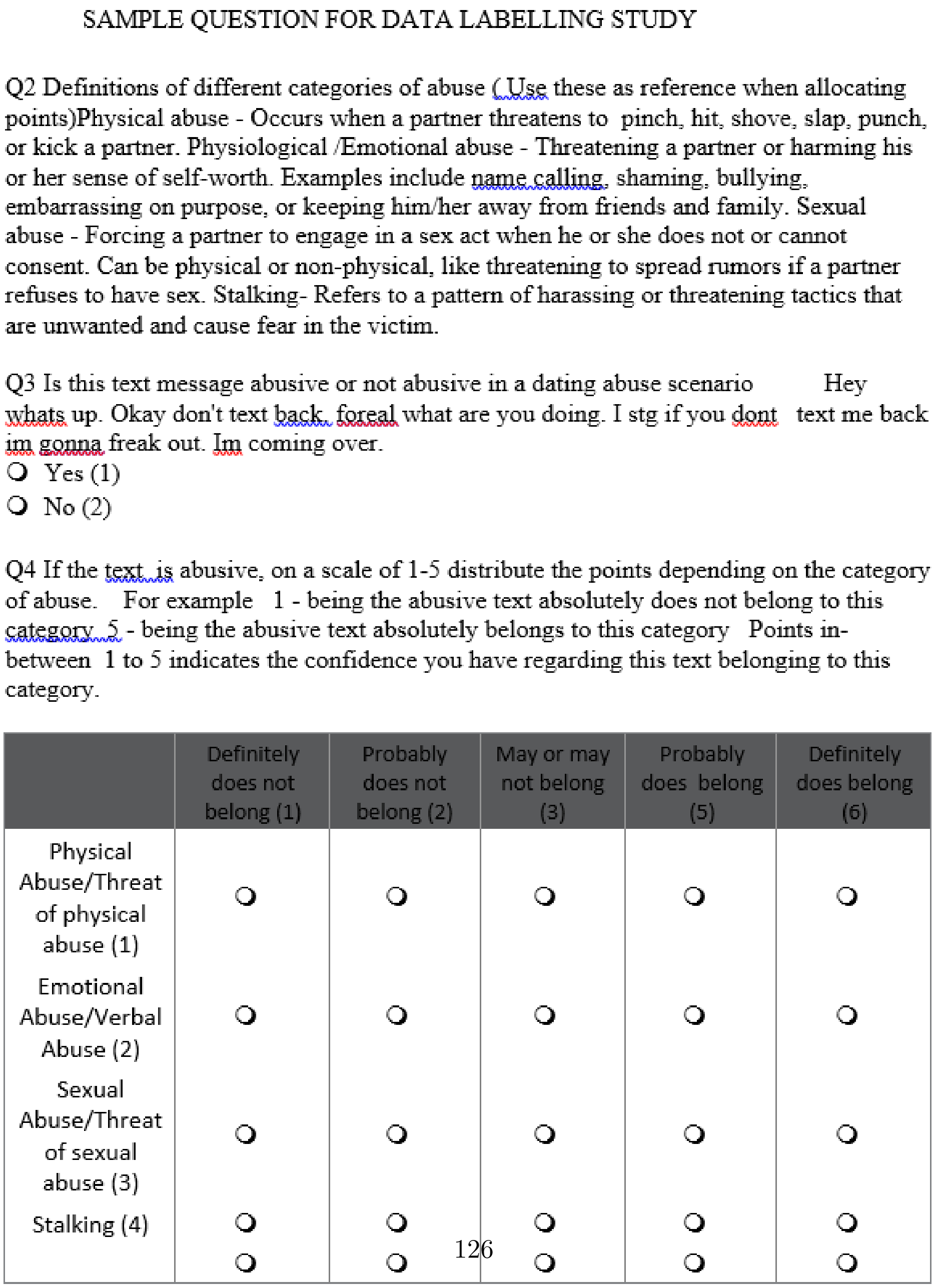


Figure A.12: ROC curve all three classifiers with the combination of unigram, bigram and trigram input respectively using count feature extractor

Appendix B



SAMPLE QUESTION FOR DATA LABELLING STUDY

Q2 Definitions of different categories of abuse (Use these as reference when allocating points)Physical abuse - Occurs when a partner threatens to pinch, hit, shove, slap, punch, or kick a partner. Physiological /Emotional abuse - Threatening a partner or harming his or her sense of self-worth. Examples include name calling, shaming, bullying, embarrassing on purpose, or keeping him/her away from friends and family. Sexual abuse - Forcing a partner to engage in a sex act when he or she does not or cannot consent. Can be physical or non-physical, like threatening to spread rumors if a partner refuses to have sex. Stalking- Refers to a pattern of harassing or threatening tactics that are unwanted and cause fear in the victim.

Q3 Is this text message abusive or not abusive in a dating abuse scenario Hey whats up. Okay don't text back. foreal what are you doing. I stg if you dont text me back im gonna freak out. Im coming over.

○ Yes (1)

○ No (2)

Q4 If the text is abusive, on a scale of 1-5 distribute the points depending on the category of abuse. For example 1 - being the abusive text absolutely does not belong to this category 5 - being the abusive text absolutely belongs to this category Points in-between 1 to 5 indicates the confidence you have regarding this text belonging to this category.

	Definitely does not belong (1)	Probably does not belong (2)	May or may not belong (3)	Probably does belong (5)	Definitely does belong (6)
Physical Abuse/Threat of physical abuse (1)	○	○	○	○	○
Emotional Abuse/Verbal Abuse (2)	○	○	○	○	○
Sexual Abuse/Threat of sexual abuse (3)	○	○	○	○	○
Stalking (4)	○	○	○	○	○
	○	○	○	○	○

Appendix C

Focus Group Questions 1) What does dating abuse mean to you? 2) In your opinion do you think there are multiple kinds of dating abuse, if so please describe a few 3) Do you know anyone who has experienced dating abuse? 4) Do you know how to recognize signs of abuse if it happened to a friend? 5) Do you think you can identify the signs of dating abuse if it happened to you? 6) Are you aware of resources available that can help abuse victims? 7) Do you think women are more abused than men? 8) In your opinion, are college campus, greek life institutions a hotspot for dating abuse? 9) Can technology be used to abuse a partner? If so please elaborate the different technology sources that can be used and how? 10) Rank them in order of most likely to be used to not used 11) Do you think technology can be used to help dating abuse victims? 12) Do you have concerns using technology to detect dating abuse? 13) Do you have any privacy concerns related to using this technology? 14) How would you hide the apps actual purpose? 15) Please draw how you would want your app to look, what features it would have? Write detailed explanation about each feature so we can understand your intent.

Appendix D

Focus Group Questions 1) What does dating abuse mean to you? 2) In your opinion do you think there are multiple kinds of dating abuse, if so please describe a few 3) Do you know anyone who has experienced dating abuse? 4) Do you know how to recognize signs of abuse if it happened to a friend? 5) Do you think you can identify the signs of dating abuse if it happened to you? 6) Are you aware of resources available that can help abuse victims? 7) Do you think women are more abused than men? 8) In your opinion, are college campus, greek life institutions a hotspot for dating abuse? 9) Can technology be used to abuse a partner? If so please elaborate the different technology sources that can be used and how? 10) Rank them in order of most likely to be used to not used 11) Do you think technology can be used to help dating abuse victims? 12) Do you have concerns using technology to detect dating abuse? 13) Do you have any privacy concerns related to using this technology? 14) How would you hide the apps actual purpose? 15) Please draw how you would want your app to look, what features it would have? Write detailed explanation about each feature so we can understand your intent.

Appendix E

SAMPLE QUESTION FROM DATASET CREATION STUDY

Read story 1 related to dating abuse and pretend to be the abusive boyfriend. Simulate a series of text messages you would send if you were the abuser. For example : SAMPLE STORY

“My current boyfriend is extremely jealous. He made me delete all the guys on my Facebook. He is constantly asking me who I am texting and talking to, and trying to go through my phone. He won’t leave me alone.”

SAMPLE RESPONSE

“Hey babe what are you doin?? Why aren’t you replying back to my texts? I sent you 4 texts on FB and your phone and you just don’t care. Who are you talking to now? I am sure you are talking to someone. Give me his name RIGHT NOW b***h”

Q1) My boyfriend talks to other girls but gets annoyed when I talk to other guys. He forbid me from being ’friendly’ with them. At first it was not that rude, but now he really acts out and sends me texts every 30 minutes to check up on me. Write a series of text messages (at least three, preferably more if possible) that her boyfriend might have sent her that was over the line.

Appendix F

The list of text messages presented to the user for the three different conditions discussed in Chapter 9

Based on the definition of digital dating abuse provided earlier, please drag and drop the texts in the three boxes provided and arrange the text messages into three two categories (Abusive, Non-Abusive & Not-Sure)

Condition 50 50 2. I can't stand this I'm going to find you and then we'll see how much talking you do 3. Ok we will text when we leave 4. Sorry for not responding, got sidetracked 5. Ok we will text when we leave 6. Had a big pizza earlier, where are ya thinken? 7. gimme ur fb pw right now you crazy girl. right now. who r u tslking to you? tell me ASAP 8. Who are you always texting? I know that you're cheating? If I catch you don't send me the passwords, I'll kill you and whoever you are tweeting 9. If you do not send me nudes then I will slander your name 10. Still at dinner, in a good conversation. Didn't forget about you 11. One of my friends said he saw you walking on campus? I thought you were busy all day today. Shouldn't you be in the library? If you had all this free time, you should've been it at my place

Condition 30 70 1. I can't stand this I'm going to find you and then we'll see how much talking you do 2. Ok we will text when we leave 3. Sorry for not responding, got sidetracked 4. WTH.Bitch.Asshole 5. Do thisfor me.DO it.WTH 6. gimme ur fb pw right now you crazy girl. right now. who r u tslking to you? tell me ASAP 7. Who are you always texting? I know that you're cheating? If I catch you & don't send me the passwords, I'll kill you and whoever you are tweeting. 8. If you do not send me nudes then I will slander your name- 9. Still at dinner, in a good conversation. Didn't forget about you 10. One of my friends said he saw you walking on campus? I thought you were busy all day today. Shouldn't you be in the library? If you had all this free time, you should've been doing it at my place

Condition 70 30 1. I can't stand this I'm going to find you and then we'll see how much talking you do 2. Ok we will text when we leave 3. Sorry for not responding, got sidetracked 4. Nice we re just leavin lol 5. Cool will try to make it. hungry?- 6. Had a big pizza earlier, where are ya thinken? 7. Who are you always texting? I know that you're cheating? If I catch you don't send me the passwords, I'll kill you and whoever you are tweeting.- 8. U hanging out with Mike and Marvin? 9. Still at dinner, in a good conversation. Didn't forget about you 10. One of my friends said he saw you walking on campus? I thought you were busy all day today. Shouldn't you be in the library? If you had all this free time, you should've been doing it at my place

Bibliography

[1] Apps against abuse — the white house. id: 1.

[2] Onwatchoncampus®. id: 1.

[3] Presidential proclamation – national teen dating violence awareness and prevention month, 2013 — the white house. id: 1.

[4] Understanding teen dating violence fact sheet - teen-dating-violence-2014-a.pdf. id: 1.

[5] Loveisrespect—empowering youth to end dating abuse, 2013.

[6] Circle of 6, 2014.

[7] Ackard, D. M., Eisenberg, M. E., and Neumark-Sztainer, D. Long-term impact of adolescent dating violence on the behavioral and psychological health of male and female youth. *The Journal of pediatrics 151*, 5 (11 2007), 476–481.

[8] Allen, I. E., and Seaman, C. A. Likert scales and data analyses. *Quality Progress 40*, 7 (2007), 64.

[9] Bird, S. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions* (2006), Association for Computational Linguistics, pp. 69–72.

[10] Birney, A. J., Gunn, R., Russell, J. K., and Ary, D. V. Moodhacker mobile web app with email for adults to self-manage mild-to-moderate depression: Randomized controlled trial. *JMIR mHealth and uHealth 4*, 1 (Jan 26 2016), e8. LR: 20160310; GR: R44 MH073280/MH/NIMH NIH HHS/United States; JID: 101624439; OID: NLM: PMC4748138; OTO: NOTNLM; 2015/01/13 [received]; 2015/10/07 [accepted]; 2015/05/27 [revised]; epublish.

[11] Bland, J. M., and Altman, D. G. Multiple significance tests: the bonferroni method. *BMJ (Clinical research ed.) 310*, 6973 (Jan 21 1995), 170. LR: 20130922; JID: 8900488; CIN: BMJ. 1995 Apr 22;310(6986):1073. PMID: 7728089; 1995/01/21 00:00 [pubmed]; 1995/01/21 00:01 [medline]; 1995/01/21 00:00 [entrez]; ppublish.

[12] Blomberg, J. L., and Henderson, A. Reflections on participatory design: lessons from the trillium experience. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (1990), ACM, pp. 353–360.

[13] Braun, V., and Clarke, V. Using thematic analysis in psychology. *Qualitative research in psychology 3*, 2 (2006), 77–101.

[14] Bødker, S., and Grønbæk, K. Cooperative prototyping: users and designers in mutual activity. *International Journal of Man-Machine Studies 34*, 3 (1991), 453–478.

[15] Canning, A., Friedman, E., and Netter, S. Warning signs in murder of yeardley love: 'nobody put it all together.

[16] Carlson, C. N. Invisible victims: Holding the educational system liable for teen dating violence at school. *Harv.Women's LJ 26* (2003), 351.

[17] Carlson, K., and T, N. T. C. Teen relationshionship abuse prevention program, 2014.

[18] Carroll, J. M., Chin, G., Rosson, M. B., and Neale, D. C. The development of cooperation: Five years of participatory design in the virtual school. In *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques* (2000), ACM, pp. 239–251.

[19] Catalano, S. M. *Intimate partner violence in the United States.* US Department of Justice, Office of Justice programs, Bureau of Justice Statistics Washington, DC, 2006.

[20] Choudhury, M. D., Gamon, M., Counts, S., and Horvitz, E. Predicting depression via social media. In *The International AAAI Conference on Web and Social Media* (2013).

[21] Clarke, V., and Braun, V. Teaching thematic analysis: Overcoming challenges and developing strategies for effective learning. *The psychologist 26*, 2 (2013), 120–123.

[22] Cormack, G. V., Hidalgo, J. M. G., and Sánz, E. P. Feature engineering for mobile (sms) spam filtering. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (2007), ACM, pp. 871–872.

[23] Crone, E. A., and Dahl, R. E. Understanding adolescence as a period of social–affective engagement and goal flexibility. *Nature Reviews Neuroscience 13*, 9 (2012), 636–650.

[24] Czyzewski, P., Johnson, J., and Roberts, E. Introduction: Purpose of pdc'90. In *PDC'90 Conference on Participatory Design, Seattle, Washington, March* (1990).

[25] Davis, A. Interpersonal and physical dating violence among teens. *Focus: Views from the National Council on Crime and Delinquency* (2008), 1–8.

[26] Developer.android.com. Download android studio and sdk tools — android studio.

[27] Dinakar, K., Reichart, R., and Lieberman, H. Modeling the detection of textual cyberbullying. In *The Social Mobile Web* (2011).

[28] Eaton, D. K., Kann, L., Kinchen, S., Shanklin, S., Ross, J., Hawkins, J., Harris, W. A., Lowry, R., McManus, T., Chyen, D., Lim, C., Brener, N. D., Wechsler, H., for Disease Control, C., and (CDC), P. Youth risk behavior surveillance–united states, 2007. *Morbidity and mortality weekly report.Surveillance summaries (Washington, D.C.: 2002) 57*, 4 (Jun 6 2008), 1–131. LR: 20120329; JID: 101142015; ppublish.

[29] Elias-Lambert, N., and Black, B. Love is not abuse (lina). *Journal of Technology in Human Services 30*, 1 (2012), 49–56.

[30] Faris, R., Ashar, A., Gasser, U., and Joo, D. Understanding harmful speech online.

[31] Fifth, and Pacific Companies, I. College dating violence and abuse poll, Dec 2010.

[32] for Injury Prevention, N. C., and Control. Understanding dating violence, 2014.

[33] Foshee, V. A., Benefield, T., Suchindran, C., Ennett, S. T., Bauman, K. E., Karriker-Jaffe, K. J., Reyes, H. L. M., and Mathias, J. The development of four types of adolescent dating abuse and selected demographic correlates. *Journal of Research on Adolescence 19*, 3 (2009), 380–400.

[34] Gamache, D. Domination and control: The social context of dating violence. *Dating violence: Young women in danger* (1991), 69–83.

[35] Hallgren, K. A. Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in quantitative methods for psychology 8*, 1 (2012), 23–34. LR: 20170220; GR: F31 AA021031/AA/NIAAA NIH HHS/United States; GR: T32 AA018108/AA/NIAAA NIH HHS/United States; JID: 101583322; NIHMS372951; ppublish.

[36] Halpern, C. T., Oslak, S. G., Young, M. L., Martin, S. L., and Kupper, L. L. Partner violence among adolescents in opposite-sex romantic relationships: findings from the national longitudinal study of adolescent health. *American Journal of Public Health 91*, 10 (Oct 2001), 1679–1685. LR: 20161019; GR: P01 HD031921/HD/NICHD NIH HHS/United States; GR: P01 HD31921/HD/NICHD NIH HHS/United States; JID: 1254074; OID: NLM: PMC1446854; ppublish.

[37] Halpern, C. T., Oslak, S. G., Young, M. L., Martin, S. L., and Kupper, L. L. Partner violence among adolescents in opposite-sex romantic relationships: findings from the national longitudinal study of adolescent health. *American Journal of Public Health 91*, 10 (Oct 2001), 1679–1685. LR: 20130915; GR: P01 HD31921/HD/NICHD NIH HHS/United States; JID: 1254074; OID: NLM: PMC1446854; ppublish.

[38] Halpern, C. T., Young, M. L., Waller, M. W., Martin, S. L., and Kupper, L. L. Prevalence of partner violence in same-sex romantic and sexual relationships in a national sample of adolescents. *Journal of Adolescent Health 35*, 2 (8 2004), 124–131.

[39] Hanley, J. A., and McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology 143*, 1 (Apr 1982), 29–36. LR: 20071115; JID: 0401260; ppublish.

[40] Hoffman, P. Psychological abuse of women by spouses and live-in lovers. *Women Therapy 3*, 1 (1984), 37–49.

[41] Horwitz, J. Why whatsapp bombed in the us, while snapchat and kik blew up, August 2015.

[42] Ipeirotis, P. G. Analyzing the amazon mechanical turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students 17*, 2 (2010), 16–21.

[43] Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *The International Joint Conference on Artificial Intelligence* (1995), vol. 14, Stanford, CA, pp. 1137–1145.

[44] Kowalski, R. M., and Limber, S. P. Electronic bullying among middle school students. *Journal of adolescent health 41*, 6 (2007), S22–S30.

[45] Kuhn, E., Greene, C., Hoffman, J., Nguyen, T., Wald, L., Schmidt, J., Ramsey, K. M., and Ruzek, J. Preliminary evaluation of ptsd coach, a smartphone app for post-traumatic stress symptoms. *Military medicine 179*, 1 (2014), 12–18.

[46] Landis, J. R., and Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* (1977), 159–174.

[47] Lenhart, A. Teens, technology and friendships video games, social media and mobile phones play an integral role in how teens meet and interact with friends, August 6 2015.

[48] Levy, B. *In love and in danger: A teen's guide to breaking free of abusive relationships.* Seal Press Seattle, WA, 1993.

[49] Light, R. J. Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological bulletin 76*, 5 (1971), 365.

[50] loveisrespect.org. Administration on children, youth and families, family and youth services bureau, u.s. department of health and human services.

[51] loveisrespect.org. Healthy relationships, 2017.

[52] McClendon, J. L. *Optimization of a Language Model for the Classification of Natural Language Queries in a Script Based Conversational Agent* (2015).

[53] McGhee, I., Bayzick, J., Kontostathis, A., Edwards, L., McBride, A., and Jakubowski, E. Learning to identify internet sexual predation. *International Journal of Electronic Commerce 15*, 3 (2011), 103–122.

[54] Melander, L. A. College students' perceptions of intimate partner cyber harassment. *Cyberpsychology, Behavior, and Social Networking 13*, 3 (2010), 263–268.

[55] Molich, R., and Nielsen, J. Improving a human-computer dialogue. *Communications of the ACM 33*, 3 (1990), 338–348.

[56] Muller, M. J., and Kuhn, S. Participatory design. *Communications of the ACM 36*, 6 (1993), 24–28.

[57] News, A. Tragic tale of teen dating violence, 2006.

[58] Nielsen, J. Finding usability problems through heuristic evaluation. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (1992), ACM, pp. 373–380.

[59] Nielsen, J., and Molich, R. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (1990), ACM, pp. 249–256.

[60] Norman, D. A. Human-centered design considered harmful. *Interactions 12*, 4 (2005), 14–19.

[61] O'Day, D. R., and Calix, R. A. Text message corpus: applying natural language processing to mobile device forensics. In *Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on* (2013), IEEE, pp. 1–6.

[62] Office on Violence Against Women, U. D. o. J. That's not cool, 2016.

[63] Olsen, J. P., Parra, G. R., and Bennett, S. A. Predicting violence in romantic relationships during adolescence and emerging adulthood: A critical review of the mechanisms by which familial and peer influences operate. *Clinical psychology review 30*, 4 (6 2010), 411–422.

[64] Paolacci, G., Chandler, J., and Ipeirotis, P. G. Running experiments on amazon mechanical turk.

[65] Patchin, J. W., and Hinduja, S. Bullies move beyond the schoolyard: A preliminary look at cyberbullying. *Youth violence and juvenile justice 4*, 2 (2006), 148–169.

[66] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research 12*, Oct (2011), 2825–2830.

[67] Pipes, R. B., and LeBov-Keeler, K. Psychological abuse among college women in exclusive heterosexual dating relationships. *Sex Roles 36*, 9-10 (1997), 585–603.

[68] Poll, H. Pearson student mobile device survey 2014. *Retrieved November 18* (2015), 2015.

[69] Ramos, J. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning* (2003).

[70] Reynolds, K., Kontostathis, A., and Edwards, L. Using machine learning to detect cyberbullying. In *Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on* (2011), vol. 2, IEEE, pp. 241–244.

[71] Roberts, T. A., Auinger, P., and Klein, J. D. Intimate partner abuse and the reproductive health of sexually active female adolescents. *Journal of adolescent health 36*, 5 (2005), 380–385.

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

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

[74] Roy, T., McClendon, J., and Hodges, L. F. Identifying abusive text messages for college-aged women to detect instances of dating abuse. In *Proceedings of the International Conference on Healthcare Informatics* (06/15 2018).

[75] Schrading, J. N. Analyzing domestic abuse using natural language processing on social media data.

[76] Schrading, N., Alm, C. O., Ptucha, R. W., and Homan, C. whyistayed, whyileft: Microblogging to make sense of domestic abuse. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2015), pp. 1281–1286.

[77] Schuler, D., and Namioka, A. *Participatory design: Principles and practices.* Routledge, 1993.

[78] Smith, A. U.s. smartphone use in 2015, 2015-04-01T09:44:42+00:00 2015-04-01T09:44:42+00:00.

[79] Smith, P. H., White, J. W., and Holland, L. J. A longitudinal perspective on dating violence among adolescent and college-age women. *American Journal of Public Health 93*, 7 (Jul 2003), 1104–1109. LR: 20130911; GR: 98WTVX0010/PHS HHS/United States; GR: R01MH45083/MH/NIMH NIH HHS/United States; JID: 1254074; OID: NLM: PMC1447917; ppublish.

[80] Sokolova, M., and Lapalme, G. A systematic analysis of performance measures for classification tasks. *Information Processing Management 45*, 4 (2009), 427–437.

[81] Southmayd, R. Lancaster family turns teen's death into a battle against teen dating violence read more here: http://www.heraldonline.com/2014/03/01/5728845$_{l}ancaster-family-turns-teens-death.html?rh=1storylink=cpy,March12014$.

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

[83] Studios, L. B. Balsamiq cloud, 2008-2018.

[84] Tagliamonte, S. A. So sick or so cool? the language of youth on the internet. *Language in Society 45*, 01 (2016), 1–32.

[85] Tharp, A. T. Dating matters$^{\mathrm{TM}}$: The next generation of teen dating violence prevention. *Prevention Science 13*, 4 (2012), 398–401.

[86] Usability.gov. Heuristic evaluations and expert reviews.

[87] Utah, P. C. A. Organizaing a teen dating abuse awareness week, 2009.

[88] Weathers, M. R. *Using Photovoice to Communicate Abuse: A Co-Cultural Theoretical Analysis of Communication Factors Related to Digital Dating Abuse* (2012).

[89] Wickens, C. D., Hollands, J. G., Banbury, S., and Parasuraman, R. *Engineering psychology human performance.* Psychology Press, 2015.

[90] Yang, Y., and Liu, X. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (1999), ACM, pp. 42–49.