Detecting Abnormal Social Robot Behavior through Emotion Recognition

Subhash Rajapaksha Rajapaksha Pathiranage
Marquette University

DETECTING ABNORMAL SOCIAL ROBOT BEHAVIOR THROUGH EMOTION RECOGNITION

by

Rajapaksha Pathiranage, Subhash

A Thesis submitted to the Faculty of the Graduate School,
Marquette University,
in Partial Fulfillment of the Requirements for
the Degree of Master of Science

Milwaukee, Wisconsin

May 2019

ABSTRACT
DETECTING ABNORMAL SOCIAL ROBOT BEHAVIOR THROUGH
EMOTION RECOGNITION

Rajapaksha Pathiranage, Subhash

Marquette University, 2019

Sharing characteristics with both the Internet of Things and the Cyber Physical Systems categories, a new type of device has arrived to claim a third category and raise its very own privacy concerns. Social robots are in the market asking consumers to become part of their daily routine and interactions. Ranging in the level and method of communication with the users, all social robots are able to collect, share and analyze a great variety and large volume of personal data.

In this thesis, we focus the community's attention to this emerging area of interest for privacy and security research. We discuss the likely privacy issues, comment on current defense mechanisms that are applicable to this new category of devices, outline new forms of attack that are made possible through social robots, highlight paths that research on consumer perceptions could follow, and propose a system for detecting abnormal social robot behavior based on emotion detection.

ACKNOWLEDGEMENTS

Rajapaksha Pathiranage, Subhash

This thesis was accomplished with continuous support from my advisor, thesis committee members (Drs. Kaczmarek and Madiraju), and my family. In addition, this research is funded by National Science Foundation (NSF) under research project entitled as “CRII: SaTC: Towards Securing Social Robots” and Northwestern Mutual Data Science Institute Student Scholarship (NM DSI Scholarship) for spring 2019.

First, I want to thank Dr. Debbie Perouli, my Advisor, Director of Cyber Security Lab, for all the suggestions and guidance she gave me. She guided the direction of my research, helped me implement applications, and inspired me to confront all the difficulties and troubles during the research and thesis writing. I would like to thank my wife for all the support, encouragement and sharing the burden with me. Finally, I am very grateful to my parents for their support and encouragement while I was studying, researching, and all the great things they did for me throughout my life.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

The science fiction universe in which humanoids can outperform humans in mostly every important aspect of life is still far from reality. However, our privacy is challenged in the present time by robots offering to socially engage with us. Unlike a personal digital assistant (e.g. Amazon Alexa, Google Home, Apple Siri, and Microsoft Cortana) whose main task is to correctly answer questions leveraging a wealth of publicly available data and private information accumulated per user over a period of time, a social robot's goal is to provide meaningful social interactions. Depending on the robot's design, these interactions may be facilitated through verbal cues.

Jibo does not have to wait for a user prompt, but it can initiate a discussion. Jibo also shows its interest in the human in the room by turning its "face" towards the origin of major sounds that could signify voice or movement. More advanced in its features and approximately twenty times more expensive, Softbank Pepper's top priority is to "perceive emotions". While these are some examples of robots currently in the market or soon to be released, a significant number of startups and major companies are investing their resources in their own versions of a social robot (SR).

Evolution of technologies such as artificial intelligence and data sciences are playing a significant role in every industry today. Various systems are established in collecting data, processing and functioning based on these emerging technologies. Cyber Physical Systems (SPC) and millions of Internet of Things (IoT) devices provide various services worldwide. Concepts such as smart homes have opened doors to bring more devices to the house that are connected to the network. Therefore, within the computer science community, the rise of the SR as a product does not come as a radical surprise. Leveraging advancements in the smart phone technology, a SR draws characteristics from the Internet of Things (IoT) category. On one hand, it shares limitations on resources such as computing power; on the other hand, the SR's complete functionality often relies on a connection to the manufacturer's cloud for services like face recognition. At the same time, due to its actions on the physical domain, a SR can also be categorized as a Cyber Physical System (CPS). Therefore, research from both IoT and CPS areas together with research on privacy and security related to traditional computing devices needs to be examined in order to identify both existing privacy preserving mechanisms and research areas where growth is necessary.

Our goal is to bring to the spotlight this emerging area of concern for personal privacy as SRs have started entering households and facilities that provide care to patients or older adults. In this

paper, a) we describe the risks associated with the vast availability, in both quality and quantity, of personal data to a SR that is collecting and inferring information about people within its proximity; b) we briefly survey current solutions that stem from research in authentication, encryption, CPS and intrusion detection, in particular with relation to rootkits and IoT botnets; c) we outline new forms of attacks that could materialize through a SR; d) we highlight research directions for exploration in the context of consumer perceptions in this new environment; and e) we propose a system for detecting SR misbehavior based on emotion detection.

1.1 Personal Data at Stake

The hardware components that are present in many SRs include one or more cameras, microphones, temperature and motion sensors. In addition, almost all robots have some degree of movement available, which ranges from being stationary, but able to face different directions, to complete mobility. The freedom of movement and the means through which everyday life events can be recorded in a household by a computing device is unprecedented.

The advent of the IoT and smart home devices that collect data across the house about a variety of use patterns (e.g. efficient light usage or temperature control) has already brought the issue of privacy to the forefront. If a SR is present in a house that has IoT devices, it is likely that the robot will act as the central controller of the IoT devices: the IoT microcontrollers could be sending updates to the robot on actions that the homeowner is advised to take or just notifications on the IoT device's operation. The robot will be able to communicate the messages from the IoT device in a more social or effective way than the IoT device. For instance, in the event that a water sensor detects flooding in a basement, the IoT microcontroller could notify both the owner's cellphone through a text message and the SR. At night, the owner may ignore the text message, but the robot could look for the owner in the house and make sure to communicate the event, e.g. by waking him/her up.

Since more and more in-home medical devices become available, a medical IoT hub needs to collect and combine the information from all of them. If the robot acts as the hub, then it could also be in possession of sensitive medical records.

Apart from the data that the robot collects through its own sensors and the data that IoT devices may be sharing, the SR is very likely to learn even more, including: questions that the user cares about, such as web queries, the user's food, music and fashion tastes, routine patterns like the times the user is away from home, the number of people living in the same household and who they

are, how the children of the house look like and how their voices sound, the extent to which the tastes of the various household members match. What is more, most SRs run machine-learning algorithms that will allow them to infer even information about their users. As a result, the wealth of data that the SRs are expected to access, in both quantity and quality, is almost unlimited.

Concerning is not only the profound data access that a SR will have, but also the inherent limitations the users of the robot may have in realizing or effectively preventing privacy breaches. The ease of use promise and the social interaction features make the robot a strong candidate as a child's playmate, a recuperating individual's assistant or an older adult's companion. Reading bedtime stories and reminding someone to take medical pills at the prescribed times are among the use cases that have been advertised. If sections of the population that are by default more vulnerable to privacy attacks, e.g. due to limited exposure to technology, are going to be the main users of SRs, then the level of concern about protecting the privacy of those people should be even higher.

1.2 Consumer Perception

The number of SRs already available for sale or advertised to reach the public within months has steadily increased in the United States during the last couple of years. At present, there are about twenty such robots, while the number is likely to be higher in other countries, such as Japan. The market research seems to imply that consumers are likely to adopt this new type of computing device in their homes or workplaces. In this set of new products, one can also add the personal, voice-activated assistants that are already successful but currently lack mobility and some of the more advanced social skills. The welcoming of personal assistants is accredited to a large extent to the perception of the consumers, if not reality, that these devices are easier to use than a smartphone or a tablet for certain tasks. The promise of home robots is that they will increase not only usability, but also the satisfaction one gets from interacting socially with artificial intelligence.

We believe that it is valuable to explore three research directions that relate to consumer attitudes and security in this new environment. One potential danger for consumers is to perceive SRs as more secure than other computing devices, if the robots succeed in presenting themselves as likable and trustworthy. Science fiction and the media have been presenting for decades ideas to the public about the dangers of artificial intelligence. However, the notion that a robot may sound more intelligent than a laptop, but could be more vulnerable to specific security attacks (e.g. due to less frequent patching) has not enjoyed similar attention. Second, we should investigate whether the SR

manufacturers take steps to make the robots not only easier to use, but also easier to secure. A device that is capable of satisfying user needs with minimal human intervention can create the false impression of being able to maintain its security protections without significant user participation. If a SR's security depends on user action as much as securing a desktop, it is reasonable to expect that consumers will be deceived, even unintentionally, regarding the robot's security.

Finally, it is worth noting that for a portion of the market of SRs, the individual making the financial investment to buy it and the individual using the robot could be different. The SRs are advertised to appeal among others to parents, so that they purchase the robot to entertain their children, and to adult children who wish to aid their own parents with common older age problems, such as weak memory and loneliness. As it has happened with personal assistants, these SRs could also be presented as a gift to someone else. If there is a significant percentage of purchases where the person deciding to buy the robot is different from the person using the robot, then it is worth asking whether the attitude of the user regarding the privacy guarantees of the product depends on who made the financial investment.

2 SURVEY OF CYBER-ATTACKS ON SOCIAL ROBOTS

The computing capabilities of a SR tend to lie somewhere between those of an IoT device and a personal computer. Several SRs have operating systems, such as the Robot Operating System (ROS) or proprietary, that are often more lightweight than a typical Linux distribution. The memory capacity is limited, but the user can usually access cloud storage, since the robot often allows Bluetooth or WiFi connections. Battery life could be a concern and applications that require extensive computing resources are designed to run on the cloud or on a different personal device of the user.

An important question is whether current security mechanisms can be applied to the SR. The question is not merely answered by looking at the technical requirements of existing security solutions, due to the SR's central promise: it has to be easy and pleasant to use. Even transferring the standard use of passwords for authentication in the SR environment is not trivial, since many of these devices are supposed to be voice or motion triggered, work with several users, and may even lack a touch screen to receive user input. It remains to be studied how effective are the authentication mechanisms that different manufacturers have opted for, and if they have included them in the robot design in the first place.

In the following, we provide examples of defense approaches that are applicable in the SR setting against specific security attacks. We also highlight limitations of these solutions due to the inherent characteristics of a SR. We have included research studies that target vulnerabilities in SRs, articles that come from the area of cyber physical systems, approaches against botnets, and rootkit prevention and detection methods. The reason we focused on the last two types of attacks is that botnets have recently caused major concern among the IoT world (e.g. Mirai malware), while rootkits are one of the most persistent types of malware.

2.1 Survey 1: Attack Methods and Social Robot

We surveyed more than 45 security threats and attacks; those attack surfaces are core network, network edge, home network or WIFI, and in side devices. In this survey, we focused on behavior of the attack, existing protection mechanisms, and its relativity to social robots. These attacks focus on confidentiality, integrity and availability of the devices and systems. Denial of Service (Dos) attacks has different versions based on its behavior such as Distributed DoS, SYN flood, UDP flood, and overwhelming memory. This attack comes under availability, because main purpose of it is to shut

down the system. Eavesdropping, scanning and probing, packet sniffing tries to access to data such as passwords, financial data, system information, or used in monitoring system behavior. This affects the confidentiality of the system or the device. In addition to that, malware comes in various versions. They change system data and functions that affect system integrity.

There are various tools and techniques to protect systems and devices from attacks. These prevention tools act in different locations of the system and in different scales. Firewalls are one example that provide detection and protection services for the network. Antivirus applications are common prevention tools for personal devices. Anti-phishing toolbars in browsers act in a location where device or user interface with the network. In addition to that, researches are being done to invent new methods and improve accuracy of existing methods. Rootkit detection and prevention researches are a good example. That will be discussed later in this chapter. Even though these attacks commonly make impact on many of the devices and systems, the significance of the impact is different from device to device. At the same time, applicability of these prevention methods also changes based on the device. Therefore, in this survey, we tried to identify most impactful existing attacks that can affect social robots. Table 1 has summarized results of the survey that shows the most impactful existing attacks for social robots.

Table 1: Current Threats and Protections for Social Robots

Threat	Current Protection Tools and Techniques	Possible Relations to Social Robots (SR)	Limitations to use in SR
Social Engineering	Two-factor authentication, social engineering prevention methods, Anti-Phishing toolbars which compare visiting sites with phishing sites, Spam filtering	Most of the phishing attacks are based upon social engineering. In SR setting, some of them are obsolete, and some are beyond the control of SR. Ex: robots can identify visually similar but fake URLs while robots will not identify fake emails. Further, if attacker could route SR to a fake web page by other mean like setting up a temporary DNS, SR cannot identify fake visuals on that web site.	Incompatibility of prevention techniques for social engineering like making web pages personally recognizable, user awareness to use human instinct. Browser based solutions are not supporting since SR will not have browsers.

Denial of Service (DoS)	Traffic management, load balancing tools, Check point firewall, Collection of reverse proxies, Null-Routing by ISP, Router configs: Shutting Broadcast, not responding to ICMP requests, Response Rate Limiting(RRL)	SR may provide set of essential services via internet or in-house network. Ex: Online home monitoring, in-house device control. However, attack can shutdown processes or entire device, which provide essential services. Or else it can isolate the device from the network.	Limited resources in SR and personal usages at home prevent using bulky systems such as, Load balancers, firewalls, traffic management systems, reverse proxies. Network device configurations like RRL cannot be expected from users.
Botnet	Anti-social Engineering based identification and prevention methods Anti-virus software firewalls Malware detectors Honeypot/ Honeynet, IRC tracking, DNS tracking	SR can become a zombie robot. Its setup support botnet activities. There can be many idle robots to be utilized in the long-run. On the other hand, robot performance can be degraded because of botnet activities. These activities may focus on external targets, ex: Mirai.	Hard to implement prevention tools inside SR since they are light weighted and mobile. In addition, SR are not working in heavily secured sophisticated networks. Therefore, Honeypot kind of solutions are also helpless.
Rootkit	Anti-social engineering concepts for users like not clicking on unknown emails, attachments, links, installing certified software, etc. User based or Kernel based anti-rootkit software.	Different robots have different capabilities. Therefore, intruders with root access may get different capabilities. Higher capabilities higher the risk. Ex: A robot who can move around the house and pick door locks can let the intruders physically come in. A robot who has cameras, microphones may let intruders to spy.	Offline rootkit detector is not a proper solution since SR need real time solutions. Limitations in prevention methods to social engineering attacks. On the other hand bringing them in to a separate central system like a cloud for monitoring may create complex infrastructure requirements with additional heavy processes.
Zero-day Attack	Updating/Patching	SR can have this type of attacks as other systems. Moreover, in the similar way, designers have to foresee such vulnerabilities before attackers, and take necessary actions.	Knowledge of users of SR to do updates and patching
OS Vulnerabilities	Regular software patching	OS vulnerability attacks are acting in a similar way for SR as well. But since SR of should support more hardware (Motors, sensors) than a computer, it could have more vulnerability chances than a general computer	Cannot expect user to install new patches timely, because the intended users may range from small children to old patients

Social robots are vulnerable to attacks such as Spoofing, Probing, Session Hijacking, Brute force, or Dictionary. However, prevention techniques existing today for those attacks can moistly prevent social robots. Apart from that, we identified Social Engineering, DoS, Rootkits, Botnets, Zero-day attacks and OS vulnerabilities that are some of the most impactful attacks to social robots. Even though, there are security tools and techniques exist for preventing systems and devices from them, in the social robot setting, these attacks may sneak through these security tools.

Social Engineering is a very common way to get into systems or take information out. Stealing passwords, financial details (credit card numbers) and other personal information or installing malicious programs into systems are some its purposes. Natural human instinct is a key prevention method for that. Therefore, user awareness is crucial. Sheng et al[1] has done a demographic analysis of phishing susceptibility and effectiveness of interventions. In their survey, a well-designed and effective training sessions with readings, games, cartoons and web based training tools have resulted that education materials reduced 40% of tendency getting into phishing scams. In online banking, websites uses secondary questions, familiar images, nicknames and many other ways to make the page familiar to the user. However, Social robots have extended its capabilities to a level where previously users required doing. This adds an intermediate level to the user-machine interface. Hence, the user may not need to perform functions like reading emails, filling web forms or clicking URLs anymore, when social robots perform such functions instead of a human. For example, a browser-like program in the social robot may read the HTML code of web pages or a mailbox that may read emails to the user. Therefore existing user oriented prevention mechanisms will not defend social engineering based attacks with social robots any more.

Social robots may be used for security monitoring and access controlling, nursing old people, differently abled people or children, first aiding with CPR, etc. Therefore, such robots must provide a reliable service. If not, the consequences could be even life threatening. Therefore, targeting such devices in executing attacks such as DOS and botnets that stops the device from performing is serious. Firewalls, traffic management, load balancing, response rate limiting, Intrusion Prevention System (IPS) and Intrusion Detection System (IDS) are some of the protection mechanisms which are heavily used in sophisticated systems today. Kim, W. et al [2] and Carrow EL [3] provide two types of solutions for botnets. One type is common mitigation procedures such as system patch updates, disabling JavaScripts, filtering attack signatures, monitor traffic flow. Usually social robots are

lightweight and compacted. Therefore, they may not be able to facilitate for some of those protection mechanisms. The second type is enterprise level solutions such as Black Hole networks, Honeypots, IRC and DNS tracking. However, enterprise level solutions will not also support the SR with the design and setting.

Rootkit has the capability of hide itself from being detected and being inactive until the opportunity comes. Musavi SA. et al [4] says that, the rootkit uses different mechanisms such as file masquerading, redirecting execution path by hooking, direct kernel object manipulation, changing boot sequence, etc. These techniques vary from rootkit to rootkit, making the attack model more complex. Further to the paper, underground market is offering rootkit modules included in Malware-as-a-Service infrastructure. Romana S. et al [5] explains Bill Blunden's classification which classify rootkit's hooking technique based on eleven different code and data structures in user and kernel space.

Anti-rootkit software is a commonly used solution for existing systems. Among them, behavior based approach is a common in the rootkit detection. Signature based detection provides quick results. However, its accuracy is low, and it cannot identify new attack. Therefore, rootkits can hide themselves from those signatures. Behavior-based detection uses different techniques. For example, Cui W. et al [6] and Musavi SA. et al [4] use static and dynamic memory analysis in detecting rootkits. Xie X. et al [7] reconstructs the system state at hypervisor level to detect rootkits. Yin H. et al [8] uses a taint base memory access and flow monitoring technique to detect rootkits. Yin H. et al [9] and Romana S. et al [5] use resource access monitoring technique that detects hooks into libraries and OS calls. Any such detection method is good for social robots as they are quick and lightweight.

Literature provides different varieties of rootkit detecting mechanisms as well. Most of them are offline-based solutions that reconstruct the memory traces or use system images to analyze and detect rootkits. Cui W. et al [6] provides such offline memory analysis system. Nevertheless, the SR is dynamic and a real time system, which demands an online detection system. In addition to that, Virtual Machine Introspection is the currently available most effective rootkit detection technique. Some designs consist of hypervisor-based approaches such as Xie X. et al [7]. In the case of rootkits, it is interesting to note that many prevention or detection mechanisms rely on virtual memory introspection (VMI) techniques. It is questionable whether such solutions could carry over to the SR domain, since there seems to be neither good reason nor resources for supporting virtual machines on such a robot.

However, evaluation of anti-rootkit tools of Romana S. et al [5] and static memory analysis of Musavi SA. et al [4] extract features and parameters that helps in detecting rootkits. This online and lightweight approach can open up a way to the SR security model.

2.2 Survey 2: System Vulnerabilities and Social Robots

As well as looking at existing attacks types, we studied about vulnerabilities of existing social robots and systems, which are similar to social robots. Table 2 consists of a set of literature, which addresses vulnerabilities of different systems with suggested solutions. We identified the possibilities of appearing those security problems in the SR setting. For each of the system, we discuss applicability of suggested solutions to SRs and limitations for applying them.

Table 2: Existing Systems and Vulnerabilities

System	Paper	System Vulnerabilities	Protecting Techniques
SR	Denning T. et al [10], Jeong S. et al [11]	Broken Authentication of ROS, Vulnerable to ROS bag of replay attack, Vulnerabilities of ROS communication, Vulnerable to service hijacking, Remote identification & discovery, Passive & active eavesdropping, Lack of operational notification, Lack of network security	Provides a set of design questions that expose issues of social robots security, Suggest security protocols like limiting access and encryption
CPS	Templeton SJ. [12], Lin H et al [13], Junejo KN et al [14], Mitchell R. & Chen I.[15]	Poor access control, Poor input validation, Lack of robustness, Implementation errors, Limited interoperability, Lack of preventive safety, Naïve assumptions about security, Proprietary solutions, Safety lock outs	Suggestions such as awareness, standardization, certifications. Security models.
IoT	Al-Sarawi S. et al [16], Mustapha, H & Alghamdi A.M. [17], Bertino, E & Islam, N.[18] Park, J. et al[19]	Insufficient authentication/authorization, Insecure network services, Insufficient security configurability, Insecure software or firmware	Security practices such as changing default password, updating security patches, disabling Universal Plug and Play (UPnP), monitoring ports and anomalous traffic.

The logic behind selecting the Cyber Physical System (CPS) as an SR-related category is because it has a similar set of behaviors. According to Junejo KN et al [14], the CPS is defined as an overlay of cyber sensing and control over a physical system for various mission-critical tasks. Mitchell R. and Chen I. [15] look into large-scale, geographically dispersed life critical systems that comprise

sensors, actuators, controls and networking components. In comparison, The SR comprises of sensors, actuators, and networked systems operating in both cyber and physical domains, even involved in life critical operations. The cyber domain of the SR inherits from computers and the physical domain inherits from electrical, mechanical, and electronics units such as sensors, motors, etc. Therefore, the SR can also be categorized as a CPS and security related studies of CPSs work closely with SRs as well. In the same way, IoT can be considered as a kind of distributed systems of a CPS or SR. We analyzed the facts in studied literature under different sections such as vulnerabilities, attack models, detection techniques and performance measurement.

2.2.1 Vulnerabilities

The cyber domain of social robots inherits from computers. Therefore, most of the computer system vulnerabilities exist in the SR setting as well. WIFI connectivity and the internet connection can allow attackers to penetrate into the SR, application and operating system level vulnerabilities are used in attacks, or the data communication with the outside can be manipulated. In addition to that, physical domain of the SR also carries a set of vulnerabilities as well. The SR uses many sensors and actuators in the operation. Microcontrollers may not check authentication to send sensor data to the outside or they may be not using encryptions.

Some SRs have already been shown to lack fundamental security mechanisms such as proper authentication[20]. Jeong S. et al [11] discusses four vulnerabilities in the Robot Operating System (ROS), which is one of the most widely deployed operating systems in robots. The vulnerabilities include replay attacks and service hijacking, for which countermeasures are available (e.g. encryption). Giaretta et. al. [21] investigates the security levels of Pepper, a popular humanoid, and suggests improvements. Denning T. et al [10] analyzes vulnerabilities of three older household robots, Rovio, Spykee and RoboSapien, V2, such as Man in the Middle (MITM) attacks, unauthorized access to audio-visual streams and login credential leakage. The suggested solutions target the robot design phase and are structured around a set of design questions aimed at exposing privacy and security issues.

According to our study, the second system type, the CPSs include vulnerabilities that possibly appear in the SR setting. Lin H et al [13] discusses three general penetration points in the CPS design: measurement output from the hardware, measurement input to control algorithm and command input to physical process. The SR may probably has these similar penetration points. Malware can change command inputs to run a desired malicious action. Similarly, malicious activities can read and change

measurements. In addition to that, Templeton SJ. [12] talks about a set of general vulnerabilities that CPSs have such as poor access control, poor input validation, lack of robustness, implementation errors, limited interoperability, lack of prevention safety, naïve assumptions about security, etc. These also directly apply in the designing and operating phases of the SR setting.

When it comes to IoT, there are even more basic vulnerabilities exist. Mustapha, H & Alghamdi A.M. [17] identifies some of them such as Insufficient authentication and Insecure network services, Lack of transport encryption and integrity verification, Insecure software or firmware. Bertino, E & Islam, N.[18] discusses about a list of reasons for IoT security risk. Among them, IoT design issues such as not having defined perimeters, being heterogeneous with respect to communication medium, protocols, platforms and devices, which can directly be a vulnerability in social robots because social robots can be made for different purposes and usages by different manufactures, with different physical capabilities under different designs. Park, J. et al[19] also talks about similar vulnerabilities. Apart from the design, social robot and IoT have similar risky behaviors such as not existing of permission requests for installation of software and many user interactions or granular permission requests. Another common risky behavior is acting as autonomous entities that control other IoT devices.

2.2.2 Threat Models

The SR can become the target of various types of attacks. Their targets, mechanism, activating time and duration, purpose and outcome will be different from each other. Incorporating these attack models in designing prevention mechanisms is an important strategy. Concerning the social robot context study, Denning T. et al [10] discusses possible targets such as elders or children who may get damaged physically or psychologically, also mechanisms such as robot vandalism, or collective robot attacks.

In CPS study, Mitchell R. and Chen I.'s [15] survey on CPS IDS designs reflects some attack characteristics such as different durations, launching time and mechanisms, and host or network orientation. Junejo KN et al's [14] work shows a behavior of data injection or change in systems. In the same way, Lin H et al [13] and Templeton SJ. [12] discuss about targets such as data and control command integrity in CPS. Park, J. et al[19] explanations STRIDE model towards IoT that affects Confidentiality, Integrity, and Availability. The STRIDE is a model to identify security threats, which

has six components (Spoofing of user identity, Tampering Repudiation, Information disclosure (privacy breach or data leak), Denial of service (DoS), and Elevation of privilege).

As SR shares similar characteristics with CPS and IoT, it is vulnerable to most of these attack models. Their targets, purpose and outcome are different. Therefore, SR protection mechanisms need to study all the possible attack models to make it effective.

2.2.3 Detection Techniques

Unlike conventional personal computers, the SR has more facts to consider in order to securing it because of its additional functionalities and capabilities, excessive time criticality and reliability of services, possessing user's excess information and physical safety. Though the network based malicious activities can be mitigated by existing techniques, vulnerabilities inside and interacting points of SR with outside need a strong concern due to these facts. Mitchell R. and Chen I. [15] talks about four characteristics that a CPS intrusion detection should have such as Physical Process Monitoring (PPM), Closed Control Loop (CCL), Attack Sophistication (AS), Legacy Technology (LT). PPM and LT are more specific to CPS, which may come common with SRs. CCL and AS are also good aspects to be think about when designing a protection mechanism to SR. Therefore, employing intrusion detection techniques in the SR is complex than a personal computer. Further, the design of the SR is another factor that should be considered. By the design itself, some simple detection mechanisms can be originated in detecting and securing. Denning T. et al [10] says not having designed features such as generating noises when moving and stationary but active, audible alert when logging in to SRs can become a vulnerability. This is very important in human-SR interaction. Because, user awareness in one of the most important topic in the SR security.

Accuracy is another important fact that the SR detection technique needs. Low accuracy risks the safety of the user. This will lead to jeopardizing sensitive information, life critical services and physical safety. Junejo KN et al [14] evaluates machine learning (ML) classifiers on a specific CPS illustrating how privacy and security mechanisms could be enhanced through ML techniques in a time critical environments. This can be considered in developing a security mechanism to SR. Signature based and Behavior based detection approaches are common among researchers. Signature base approaches are fast, accurate and light weight but susceptible for new form of attacks while behavior based approaches are slow, low in accuracy and may be bulky. Therefore, detection technique with high accuracy is needed which has compatible characteristics with the SR setting.

Lin H et al [13] says that attacks in CPSs are difficult to detect by monitoring the cyber or physical domains separately from each other. For example, lighting intensity control command in a SR can be changed maliciously into a different legitimate value in the cyber domain, but the change happens in the physical system can harm eyes of an infant in the crib. Therefore, cyber and physical interaction and propagation of effects to sub-systems are some important facts to consider in accurate detection of malicious activities in the SR.

As above example shows SR's involvement with IoT, incorporating IoT based detection techniques are also vital. As explained in Bertino, E & Islam, N.[18], users and developers should perform known best security practices. However, SR is for different users, such as kids, patients and elder people, incorporating these practices in designing and adding additional service units for security is important.

2.2.4 Performance Metrics

Delayed detections and leaving some attacks undetected by the protection protocols makes the SR half-open to threats even when a protection mechanism is included. False alarms interrupt the services and use limited resources of SR unnecessarily. Therefore, validating the performance of detection system is critical. Mitchell R. and Chen I. [15] says that generally in detection systems false positives, false negatives and true positives are common factors to measure detecting performance, but detection latency is rarely used. Social robots are very dynamic and time-critical in behavior. A malicious change of the motor rotation speed in a control message can damage the child or the patient that the SR is serving to if the detection is delayed. Throughout the time, machines took a long time to win the trust from people for being precise and accurate due to the latency in detecting, processing and reacting. Therefore, while delays in detecting adversaries make the user vulnerable, the trust the users having towards social robots will collapse.

Mitchell R. and Chen I. [15] further says that the resource limitations, power consumption, communications overhead and processor load are also important facts in performance measuring. Social robot has extended limitations on power, memory, processing or time based on the providing service. Therefore, aforementioned factors become main conditions in measuring the performance in the SR. Junejo KN et al [14] uses a detection mechanism specific parameter (Time To Build the Model: time taken for machine learning training). Such method specific or attack model specific parameters may also be used in the performance metrics.

Therefore, identifying all the possible vulnerabilities and lining up all possible existing and expected attack models would be a good start. Based on that, designing well performing and fully compatible detection mechanism would be a vital strategy for the SR security.

2.2.5 Applicability & Limitations of Existing Solutions

Incorporating available solutions in each category to the new design of SR detection mechanism is important. As shown in the Table 2, different system types propose number of successful solutions to protect users and devices from attacks. However, it is important to discuss the applicability of them in the SR setting. Applying them into social robots depends on the technology, resources, expected outcome and compatibility of the solution to the SR.

CPSs based research papers mostly refer to a specific system to provide a solution. Lin H et al [13] discusses a behavior based detection model related to a surgical robotic system and a power grid. The model uses the incoming command and current physical state of the system to predict the system impact ahead. This is a compatible approach to the SR to develop a detection mechanism. However, social robots are not only work under commands. It has machine learning algorithms to take own decisions in certain scenarios. For example, it can respond to sounds in the environment without user involvement.

Templeton SJ. [12] gives suggestions to improve security and safety on CPSs such as education and awareness improvement for developers, standardizing security and reliability, security certifications for CPSs, safety engineering to security, product liability reforms. These are valid for most of the systems we have today including SRs. Apart from that, these literature provides information and knowledge. For example, Mitchell R. and Chen I. [15] provides informative survey with number of papers on CPSs that includes detection techniques, attack types, audit features, etc. Junejo KN et al [14] provides an evaluation of machine learning classifiers, which makes more insights to classifiers when utilizing them. The SR can employ machine learning based model to build a security algorithm. As stated in the IoT protection methods, the security practices are more essential for manufacturers than users, as they cannot be expected from users.

2.3 Unique Security Issues for Social Robots

Though OS vulnerabilities, Zero-day attacks, phishing attacks, DDoS attacks and rootkits are significant threats for social robots, they are common for many other systems. However, the social robot environment could allow older type of attacks to mutate and gain new forms.

Social engineering is a domain that is likely to be transformed. Instead of trying to create appealing electronic mail messages that can bypass spam filters, an attacker might focus on creating downloadable robot skills (the equivalent of a smart phone app) with aesthetic and socially engaging features. It is not only the financial status of a user that could be affected by attacks through social robots, but also the physical safety. In cases where the robot communicates to the user medical results or reminders, passing misinformation could result in the user suffering from medication overdose or other life threatening circumstances. If an intruder was able to access patterns that the robot has identified, the intruder would be able to strengthen any kind of attack. For example, the SR could infer the times and days that the user tends to be most tired. During these time periods, a social engineering attack will have higher probability of success. A maliciously acting robot might also pretend to place calls to friends or family members without ever attempting to actually establish the necessary network connection.

Not only mutated version of attacks, social robot setting can develop new forms of attacks as well. Since many robots have mobility, a malicious physical move against an individual with diminished mobility or low body mass could also cause significant harm. Similarly, attacks targeting a person's state of mind and mood could be developed. We categorize it to two, short-term and long-term. In short-term basis, if social interactions with an appropriately designed robot can make a person feel less isolated, tweaking the robot's behavior might result into worsening someone's mental and/or psychological status. For instance, a hacked robot might stop initiating a discussion or could suggest actions that are known to deteriorate the user's capabilities (e.g. encourage heavy drinking). In the long-term, attacks that can detect user emotion can identify user's most vulnerable time of the day and use them to push users to take attacker favorable decision. A hacked robot, which is designed to use as a teaching assistant, may be used to implant opinions on children's mind such as hate or racism.

3 USER-ROBOT INTERACTION EMOTION DETECTION

In this research, we try to concentrate on the new area, where impact on human emotion by misbehaving social robots. While SRs are becoming human-companions that could be the device that people are mostly interacting with in their daily life. This is a good opening to outsiders to monitor someone's mindset, emotions, sensitivity for things experiencing and many other mind related information. SR's also may have access to other information such as medical history, treatments, job information, financial details, online accounts, contacts, schedules, etc which may be used combined in attacks. That being said, manipulation of human emotions individually or as a group can be expected in the future. Not only the attacks, but misconfigurations, erroneous results from machine learning algorithm and many other reasons may cause similar security issues in the human robot interaction. Therefore, the importance for having solutions for these type of attacks is real as we are moving forward with these technologies.

Emotion recognition analysis is the proposing solution in this research. Therefore, we developed a feature filtering method and machine learning models for emotions recognition, which will be used in identifying emotions expressed by both the user and robot. Later, results will be analyzed. Although our focus is to identify misbehaviors, our feature filtering method is not specific to this context and can be used as a general machine learning feature selection process.

3.1 Literature Review

Emotion recognition has been researched for decades and has acquired a significant improvement collaborating with machine learning, natural language processing, audio and video signal processing. Schuller, BJ. [22] gives a good explanation for the emotion recognition road map. Researches are trying to improve emotion recognition accuracy using various combinations of features and various fusion methods. Selection of emotion states for the classification varies based upon research groups. Roh Yw et al [23] provides a list of research groups and the emotion states they used for emotion recognition. Further, it states that it is desirable to use fundamental emotions in classifications. With a background study, it continues to say that *Anger*, *Happiness*, *Sadness* and *Neutral* are the most fundamental. Categorical emotions such as *Sad*, *Happy*, *Angry*, etc are not the only outcome in emotion recognition. Aldeneh Z. and Khorram S. [24] predict *Valence* from acoustic and lexical features under several pooling options. Rong J. et al [25] classifies emotion into two basics,

negative and *positive* due to uncertainties in the definition of emotion states. Researchers use multi class classification or binary classification and fusion for combining different emotions obtained from different sources.

Speech/audio is a well-recognized source for identifying emotions. There are many researches, which focus on acoustic features. Savran, A. et al [26] uses video, audio and lexical indicators. Not only audio, video and text, Tian L. et al [27] uses dialogue cues to predict emotions. Having different sources with different feature sets, classification can be done in two ways; unimodal with combining all the features into one set or multimodal with different set of features and fusing them. Chuang, Z.J. and Wu, C.H. [28] uses a multi modal approach to classify emotion from speech and text.

Anagnostopoulos and Ilious [29] uses pitch, energy, MFCCs and Formants prosodic features and another 133 features calculated based on these four groups. Praat [30] is the tool used in feature extraction. Rong J. et al [25] uses Duration and Discrete Fourier Transformations features in addition to above four features for emotion recognition. Chen, S. et al [31] uses Continuos, Qualitative and Cepstral feature sets and statistical functions of those features for their classification. Rozgic V. et al [32] uses derived features from Mel-Frequency Cepstral Coefficients (MFCC), statistical functionals of low-level feature descriptors combining with lexical features to emotion recognition. When it comes to text feature extraction, natural language processing is a key component. Porter, M.F. [33] processes words to be used in lexical feature extraction. Jin, Q. et al [34] uses Bag of Words features and e-vector lexical features.

Feature selection is an important preprocessing step in machine learning. That helps filter out most relevant features for leaning, which increase reliability and accuracy of the modal. Li, J. et al [35] provides a detailed survey about feature selection. There are tools available for preprocessing and selecting features. Sk-learn toolkit by Pedregosa, F. [36] provides various feature selection methods such as removing feature with low variance, Univariate feature selection, Recursive feature elimination, etc. Anagnostopoulos and Ilious [29] uses WEKA[37] data mining tool for data selection. Chen, S. et al [31] does not filter features, but using all of them for classification. Yu, L. and Liu, H. [38] introduces correlation based novel concept, predominant correlation, and proposes a fast filter method which filters features without pairwise correlation analysis. Cheung, Y. and Jia, H. [39] Chormunge, S. and Jena, S. [40] discuss about feature selection using clustering.

Selection of a dataset for training emotion recognition algorithm vary from research group to group. There are available databases for emotion recognition while some groups create and use their own databases. Anagnostopoulos and Ilious [29] uses Speaker Independent Recognition in Berlin Database. AVEC2012 [41], RADVESS [42] are another popular databases for emotion recognition used by researchers.

In multi-modal classification, fusion is a process which combines the results given by the all the models. This is another topic where separate researches are taken place. Mainly there are two type of fusions; Decision level fusion which is merging decisions given by classifiers and Feature level fusion which is concatenating all the features before classification. Planet, S. and Iriondo, I. [43] is doing a comparison among these two fusion methods to identify outperforming fusion. Jin, Q. et al [34] uses different sets of classification results from different combinations of acoustic and lexical features, and use them to obtain best pair wise fusion options. Finally, fusing those best combinations to get a higher accuracy. Tian L. et al [27] introduces a Hierarchical-Level fusion strategy incorporates more abstract features.

This research focuses on predicting conversational emotions of users using utterances with emotion recognition work done before. We use multimodal binary classification, hence, prepare a foundation for developing temporal emotion patterns of the user and identifying suspicious behaviors of emotion flow in conversations between human and social robot interaction when the measuring pattern deviating from expected pattern.

3.2 Methodology

We propose a system that detects inappropriate SR behavior by keeping track of the emotions demonstrated by both the user and the SR over an interaction time window. This system is expected to be in close proximity to the user and the SR, but not part of the SR hardware configuration (e.g. separate device in the same room). Making the detection system independent of the SR adds a layer of protection in case the SR gets compromised by a malicious entity. The detection device needs to be equipped with a sensor that captures audio signals (e.g. microphone) from both the user and the SR. Alarming is a case where the system detects that the emotional condition of the user degrades after the interaction with the SR, which could be caused by either SR misconfiguration or hacking. Equally concerning could be the case where the SR demonstrates emotions that are unsuitable towards a specific user, such as anger.

Machine learning (ML) is being used in a variety of fields, including network intrusion detection, to differentiate between expected and abnormal behavior. In the domain of natural language processing (NLP), ML techniques aid in detecting variations of a person's emotions during a conversation.

We developed a multi-model detection algorithm to detect expressing emotions of both the user and the robot when they are interacting with each other. Audio is our main stream of identifying the emotion. We used filtered audio features set and trained different binary classifiers separately for each emotion. In order to identify most accurate classifier we trained eight machine-learning classifiers namely, Logistic Regression (LR), Linear Discriminant Analysis (LDA), K Neighbors Classifier (KNN), Decision Tree Classifier (CART), Gaussian NB (NB), Support Vector Classifier (SVM), Ada Boost Classifier (ABC), K Means (KMC).

We used scikit-learn [36] for training and testing of these classifiers. Scikit-learn known as sk-learn, is a commercially usable open source machine learning tool set for data mining and data analysis. We used Cross-validation in the training process, which avoids overfitting in testing. It splits the training data set into k smaller sets. Then it keeps aside one portion of sets and train the algorithm from the others. By using remaining small set, it test the trained model. This process is iteratively done until all the small sets are used as test sets. Then all results obtained in k rounds are taken into an average value, which is the accuracy of the trained model. We did this process to all the machine-learning classifiers to identify the most accurate classifier.

First, we extracted features from audio (wave files). Then we used a naval feature filtering method to identify most effective feature set for each binary classification. By using the filtered feature sets, we did binary classification to identify appearance of each emotions for the utterance. We always used the same data set for training, but when classifying one emotion all other emotions were placed in the same category. For example, when training algorithms to identify *Happiness*, all other emotions were annotated as *Other*. Then we analyzed emotion change from utterance to utterance in a dyadic conversation.

3.3 Dataset

Our experiment used The Interactive Emotional Dyadic Motion Capture database (IEMOCAP) [44], which is a multimodal, multi-speaker and acted database. This database includes

video, motion capture of face, Head Movement and Head Angle Information, audio, text data of approximately 12 hours. Ten actors have performed improvisations and scripts. Each utterance consists of categorical annotation labeling into 9 different emotion categories such as *Happiness*, *Excitement*, *Sadness*, *Anger*, *Frustration*, *Fear*, *Surprise*, *Neutral* and *Other*, as well as it consists of dimensional annotation labeling as *Activation*, *Dominance* and *Valence*. Multiple annotators have annotated each utterance and it includes a self-annotation as well. A composite categorical and dimensional value for each utterance is also included in the data set.

Out of these nine categorical emotions, we used four basic emotions namely, *Happiness*, *Sadness*, *Anger*, and *Neutral*. We considered the composite categorical annotation as the emotion representation of the utterance, since more than one annotators confirm it. The data set consists of five sessions. Each session consists of several improvisations and scripted scenarios performed by two actors. We used first three sessions of the data set for training algorithms and four and five for testing. Table 3 shows the composition of the data set.

Table 3: Dataset Used in Machine Learning

	Happiness	**Excitement**	**Sadness**	**Anger**	**Neutral**	**Total**
Training	387	504	696	606	1065	3258
Test	194	496	362	368	595	2015

3.4 Audio Analysis

In this study, we use wave files of utterances in *IEMOCAP* as audio input. *OpenSMILE* [45] is an audio feature-extracting toolkit that we utilized under the configuration according to "*The INTERSPEECH 2010 Paralinguistic Challenge*" [46]. By this configuration, we extracted 1582 features per each utterance using *OpenSMILE*.

According to *OpenSMILE* user manual document, these 1582 features includes 21 functionals (min, stddev, max, mean, etc) extracted from 34 low-level descriptors (LLD) and 34 corresponding delta coefficient ((34+34) x 21 = 1428). Then it includes 19 functionals extracted from 4 pitch-based LLD and their four delta coefficient contours (19x4x2 = 152). Finally, it includes *the number of pitch onsets* and the *total duration of the input* (2).

Further *OpenSmile* document describes that the 34 LLDs are,

- *pcm loudness* - The loudness as the normalized intensity raised to a power of 0.3
- *MFCC* - Mel-Frequency Cepstral Coefficients 0-14

- *logMelFreqBand* - logarithmic power of Mel-frequency bands 0 - 7 (distributed over a range from 0 to 8 kHz)
- *lspFreq* - The 8 line spectral pair frequencies computed from 8 LPC coefficients
- *F0finEnv* - The envelope of the smoothed fundamental frequency contour
- *voicingFinalUnclipped* - The voicing probability of the final fundamental frequency candidate. Unclipped means that it was not set to zero when is falls below the voicing threshold

In addition, the four pitch related LLDs are,

- *F0final* - The smoothed fundamental frequency contour
- *jitterLocal* - The local (frame-to-frame) Jitter (pitch period length deviations)
- *jitterDDP* - The differential frame-to-frame Jitter (the *Jitter of the Jitter*)
- *shimmerLocal* - The local (frame-to-frame) Shimmer (amplitude deviations between pitch periods)

3.4.1 Binary Classification

Binary classification refers to classifying a set into two groups. Under this method, we classified each emotion separately. We used same training data set, but when classifying one emotion all the other emotions are considered as one category. For example, when training algorithms to identify *Happiness*, all the other emotions annotated as *Other*.

Then we used a feature filtering method to identify most effective feature set for binary classification. We developed this filtering technique in order to identify most effective features in binary classification. By using this, we filtered a new feature sets for each emotion and did the binary classification for each emotion category.

For every binary classification, we trained all the selected classifiers in order to identify best performing classifier for further use. At the same time, we used cross validation in every binary classification with 10 splits and 20% validation.

3.4.2 Feature Filtering

Some features extracted from the audio, has values that can distinguish different emotion categories. Emotions such as anger and excitement have high energy than others such as sad and neutral. Each audio utterance has 1582 extracted features and some of these features may have unique

values for one emotion as opposed to others. Such features are valuable in distinguishing between emotions, while features that do not exhibit this behavior could be ignored from the classification process. In this Section, we present a new method for feature selection.

Let f_{ij} denote the value of feature i for sample j. Then, we define sets $H = \{f_{ij}$: sample has been characterized as carrying emotion Happiness$\}$ and $T = \{f_{ij}$: sample j has been characterized as carrying emotion *Sadness*, *Anger* or *Neutral*$\}$.

For each feature i in the H set we perform unsupervised K-means clustering to find the cluster center c^H_i. If f^H_{imin} is the minimum value of feature i in set H and f^H_{imax} is the maximum value of feature i in set H, then we define the range around the cluster center c^H_i as,

$$r^H_i = min\{|c^H_i - f^H_{imin}|, |c^H_i - f^H_{imax}|\}$$

Similarly, for each feature i in the T set we perform unsupervised K-means clustering to find all N cluster centers c^{Tn}_i , where $1 \le n \le N$. These cluster centers represent high-density areas of the data point distribution of other emotion categories. We can now define the distance $D^{Hn}_i = |c^H_i - c^{Tn}_i|$. If c^{Tk}_i is the cluster center for an emotion other than *Happiness* that is closest to the cluster center for *Happiness* c^H_i, then we define the minimum distance $D^H_i \equiv D^{Hk}_i$. Since different features have different data ranges, we normalize the f_{ij} values into a [0, 100] range, so that we can compare the diversity and distances among cluster centers across features.

Finally, we use the notion of inclusion to represent the percentage of cluster centers of other emotion categories that reside within the *Happiness* cluster range. We define the data sets,

$R^H_i = \{c^{Tn}_i: D^{Hn}\ i < r^H_i, 1 \le n \le N\}$ and $A^H_i = \{c^{Tn}_i, 1 \le n \le N\}$.

Then, we define inclusion as,

$L^H_i = \#R^H / \#A^H_i \times 100\%$.

It should be noted that we repeat the binary classification and derive the minimum distance and inclusion values for every emotion.

Figure 1: 'Happiness' Cluster Center Among the 'Other' Emotion Cluster Centers for the Feature pcm_loudness_sma_amean

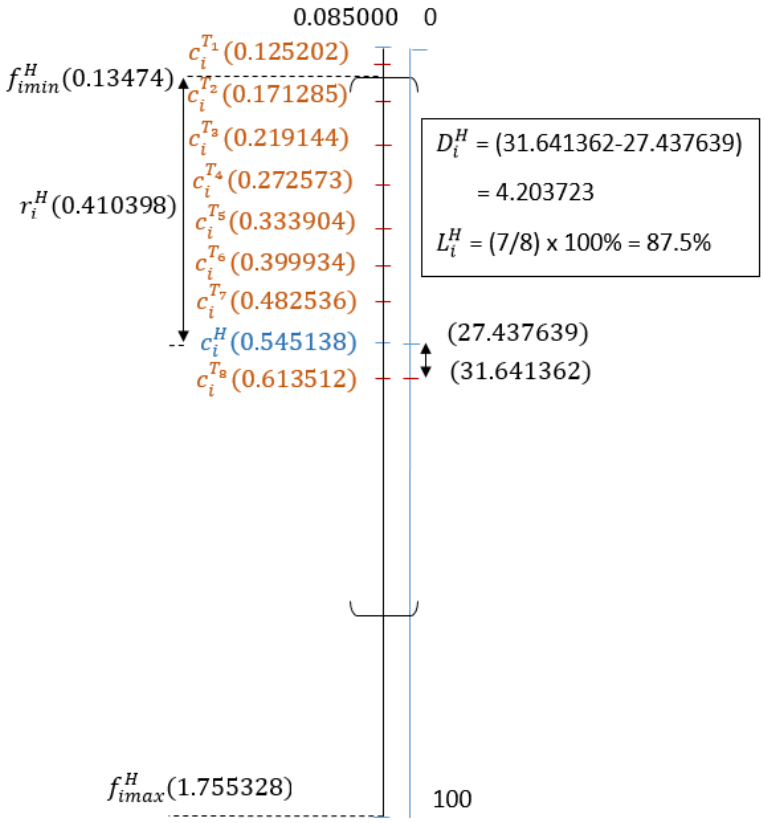


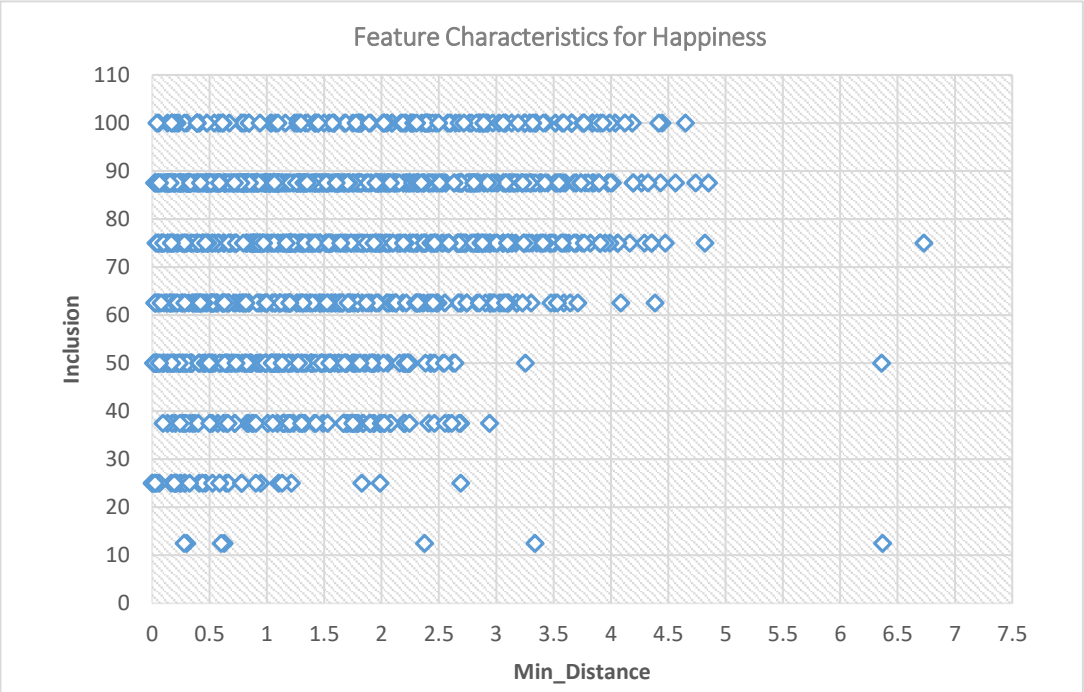
Figure 1 shows an example for feature *pcm_loudness_sma_amean* in the binary classification for emotion *Happiness*. The lowest value of this feature (0.085) was exhibited by a sample utterance that was annotated under the *Other* emotion category. The highest value of this feature (1.755328) was exhibited by a sample utterance that was annotated as *Happiness*, so this is also the f^H_{imax} value. In the normalized scale, these two values correspond to 0 and 100. For this feature, there are eight cluster centers for category *Other* and seven of those fall within the range of the *Happiness* cluster center. Therefore, inclusion is high (87.5%).

3.5 Results

3.5.1 Feature Filtering and Binary Classification

In order to identify critical features that the classification strongly depends on, we carried out the feature filtering process. After calculating *Inclusion* and *Min_Distance* for each feature for a particular emotion, we analyzed it using a graphical representation. Figure 2 shows a the *Inclusion* against *Min_Distance* for all the features for *Happiness* emotion categories.

Figure 2: Min_Distance vs. Inclusion of Features with Happiness Characteristics



The figure shows that some features do not show distinct characteristics for *Happiness* than *Other* emotion categories. Especially features that are close to (0,100) point have mixed characteristics that cannot be used to distinguish *Happiness*. Features that have higher *Min_Distance* and lower *Inclusion* explain the emotion well.

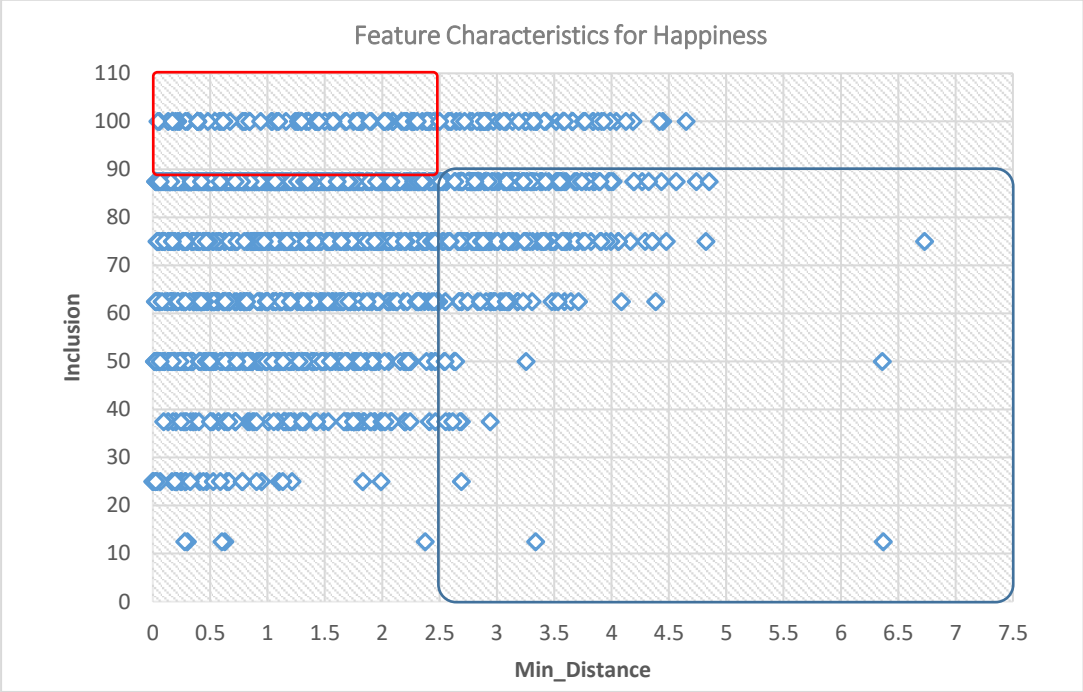
In order to confirm this, we carried out a series of tests that trains machine-learning algorithms using different combinations of these two parameters. Each combination filters outs a set of features and then those features are used to train algorithms. Table 4 shows such combinations and their results for each emotion classification. The combination notation (40 & 2) denotes *Inclusion* lower than 40% & *Min_Distance* greater than two.

Table 4: Binary Classifier Accuracy for Inclusion and Min_Distance Combinations

Combination	Anger	Happiness	Neutral	Sadness
15 & all	0.778922(LR)	0.862028(LR)	0.625051(ABC)	0.775755(ABC)
30 &_all	0.793459(ABC)	0.862028(LR)	0.649095(ABC)	0.803023(ABC)
40 &_2	0.781647(LR)	0.862028(SVM)	0.636837(ABC)	0.811174(ABC)
40 &_2.5	0.781645(LR)	0.86021(LR)	0.615045(LDA)	0.820701(SVM)
40 &_3	0.781647(LR)	0.862028(LR)	0.614591(LR)	0.814364(SVM)
50 &_2	0.863367(LR)	0.860666(LR)	0.683626(ABC)	0.826619(LR)
65 &_2	0.876991(LDA)	0.862028(SVM)	0.699556(LR)	0.828885(LR)
80 &_2	0.876991(LDA)	0.862028(SVM)	0.723137(LR)	0.82977(LR)
80 &_2.5	0.891514(LR)	0.862026(LR)	0.715411(LR)	0.833416(LDA)
80 &_3	0.891518(LR)	0.860671(LR)	0.716759(ABC)	0.830247(LR)
80 &_3.5	0.876524(LR)	0.863387(SVM)	0.699039(LDA)	0.825257(LR)
80 &_4	0.868809(LR)	0.862028(LR)	0.674101(ABC)	0.80391(LDA)
90 &_2	0.903772(LDA)	0.862028(SVM)	0.715856(LR)	0.827534(ABC)
90 &_2.5	0.900586(LDA)	0.862937(LR)	0.719484(LR)	0.838863(LR)
90 &_3	0.89832(LDA)	0.86703(LR)	0.720852(LR)	0.835699(LR)
90 &_3.5	0.889706(LR)	0.869751(LDA)	0.714486(ABC)	0.827982(LR)
100 &_2	0.887884(LDA)	0.862028(SVM)	0.719044(LR)	0.8266(LR)
all	0.875167(ABC)	0.862028(SVM)	0.705866(LR)	0.830243(ABC)

Based on this result, we obtained that the approximate highest accuracy is given at 90 & 2.5 combination. We started the test with a lower *Inclusion* value and increased it while keeping *Min_Distance at a low value*. At 90%, the accuracy got maximum and started reducing thereafter. Then we increased the *Min_Distance*. At 2.5, it gave the maximum accuracy and started reducing thereafter. Therefore, we can say that, the features that have *Inclusion* less than 90% and *Min_Distance* greater than 2.5 are most affective features. Features in the red zone in the graph affected the accuracy negatively.

Figure 3: Min_Distance VS Inclusion of Features with Happiness Characteristics - Filtered Range



We used k-means clustering for the clustering process. K-means starts clustering with arbitrary points and iteratively adjust it until it get close to center point of the data set. Therefore, for the same dataset, k-means clustering will give close but different center values in several clustering processes. Therefore, our filter method gives slightly different filtered feature sets for each clustering even when the dataset is same. However, as shown in the Table 5, results of a test that filtered same data set for hundred times, for each binary classification gave approximately similar accuracy values while the number of filtered features getting slightly changed. Figure 3 shows that there are features staying close to the selection margin. When the cluster center is slightly changing, these features are included or excluded by the margin. These features are closely located with similar characteristics. Therefore, the accuracy does not get affected significantly.

Table 5: Repeated Feature Filtering Test Results

	Classification Accuracy			**No of Features**		
	Minimum	Maximum	Average	Minimum	Maximum	Average
Happiness	0.852494	0.864296	0.858404	171	194	182.65
Sadness	0.82435	0.844306	0.835718	224	254	240.18
Anger	0.890154	0.903325	0.895928	211	240	224.78
Neutral	0.702239	0.729485	0.716392	286	313	300.13

The time taken to filter features is also another parameter to measure the performance of a filtering method. Low variance, Recursive Feature Elimination (RFE) and Univariate feature selection are feature filtering methods available in sk-learn tool [36]. We measured the average filter time of each filter and compared it with proposing filter. According to Table 6, RFE is slow compared to new feature filtering method. However, statistical base methods such as Low variance are much faster than the new method. Though Low variance is fast, it removes less amount of features. As an overall performance evaluation, new feature filtering method has considerable filter time, but selects less number of features.

Table 6: Performance Comparison with Other Common Methods

	Happiness		**Sadness**		**Anger**		**Neutral**	
	No. of Features	filter time (seconds)	No. of Features	filter time (seconds)	No. of Features	filter time (seconds)	No. of Features	filter time (seconds)
90% & 2.5	268	150.63571	356	142.67866	263	144.86095	280	128.98831
Low variance	744	0.0506901	744	0.0326299	755	0.0319149	744	0.0504839
RFE	250	705.25111	230	554.45925	260	674.56719	270	693.01954
Univariate	159	0.0662381	633	0.0682399	317	0.0785219	159	0.0667049

In order to evaluate the accuracy compared to other filtering methods, we performed binary classification from these filtered features sets by each filter method. Apart from that, we used another three feature sets used in Jin, Q. et al [34]. This features sets are also developed from same original feature set (complete set of 1582 features). These feature sets are,

- ACO - The utterance-level statistics of frame-level acoustic features, namely, continuous features (Energy: Loudness, Pitch: F0final, F0finEnv, Formants: lspFreq) and qualitative features (jitterLocal, jitterDDP, ShimmerLocal, Voicing final unclipped).
- Cepstrum - The utterance-level statistics of cepstral features (statistical functions of MFCCs (15), logMelFreqBand (8)).
- Cepstral-BoW – The bag-of-words feature representation based on frame-level cepstral features.

Both ACO and Cepstrum consist of statistics of features directly extracted from the audio. Cepstral-BoW is a set of feature set that is derived from a codebook. We generated the codebook using cepstral features. For each cepstral feature, we clustered values of each emotion category to a single point separately. For the clustering, we used session 1 and 2 of the database. From this process, we

obtained four values for each cepstral feature. Each value denotes a centroid of an emotion category. Then the codebook is generated which consists of four values for each feature that represents cluster centers for each emotion category. We created the Cepstral-BoW feature set by using this codebook. We replaced value of each feature in an utterance by the closest cluster center value in the particular feature from the codebook. To create this Cepstral-BoW feature set we used Session 2 and 3. This includes both used and unused data because session 2 is used to generate the codebook. In addition to that, we used the full feature set in this comparison by doing a binary classification using full set.

Table 7: Filtering and Classifying Comparison with Different Feature Sets

	Happiness		**Sadness**		**Anger**		**Neutral**	
	No. of Features	Accuracy	No. of Features	Accuracy	No. of Features	Accuracy	No. of Features	Accuracy
All	1582	0.862028 (SVM)	1582	0.830243 (ABC)	1582	0.875167 (ABC)	1582	0.705866 (LR)
ACO	616	0.862028 (SVM)	616	0.818885 (ABC)	616	0.86428 (ABC)	616	0.691331 (ABC)
Cepstrum	966	0.862028 (SVM)	966	0.821181 (ABC)	966	0.884245 (ABC)	966	0.71224 (LR)
Cep_BoW	966	0.864828 (ABC)	966	0.822759 (ABC)	966	0.90069 (ABC)	966	0.722069 (ABC)
90% & 2.5	268	0.862937 (LR)	356	0.838863 (LR)	263	0.900586 (LDA)	280	0.719484 (LR)
Low Variance	744	0.862028 (SVM)	744	0.826608 (ABC)	755	0.890611 (ABC)	744	0.718599 (ABC)
RFE	250	0.865662 (LDA)	230	0.842517 (LDA)	260	0.898764 (LDA)	270	0.724912 (LDA)
Univariate	159	0.867468 (LDA)	633	0.830251 (LR)	317	0.900154 (LR)	159	0.719048 (LR)

Table 7 shows the results of the tests carried out. 90% & 2.5 feature set has shown that it can give comparatively higher accuracy while having less number of features in the list. Therefore, we obtained trained binary classifiers for each emotions with higher accuracy. These classifiers take less number of inputs, which is important in a real time system for fast prediction. Using session 4 and 5 of the database, we obtained the confusion matrix as shown in Table 8.

Table 8: Confusion Matrix

	FP	**FN**	**TP**	**TN**
Happiness	0.034891376	0.11323239	0.014483213	0.837393022
Sadness	0.0520079	0.117182357	0.121132324	0.709677419
Anger	0.097432521	0.041474654	0.200789993	0.660302831
Neutral	0.171823568	0.206714944	0.184990125	0.436471363

3.5.2 Multiclass Classification

In order to compare and make sure multiclass classification results are less accurate compared to binary classification, we carried out multiclass classification. Since the new feature filtering method filters out emotion specific feature sets, it cannot be used in multiclass classification. Therefore, we used Jin, Q. et al [34] feature sets.

In multi-class classification with three feature sets, we obtained accuracies for each classifier as shown in Table 9. Logistic Regression (LR) gave maximum accuracy for Cepstrum and CepstrumBoW sets while Linear Discriminant Analysis (LDA) giving maximum accuracy for ACO. However, accuracies are in between 50-60%, which is significantly low. In this classification, we considered *Excitement* and *Happiness* as one category, expecting similar behavior.

Table 9: Multiclass Classification Accuracy of Each Classifier for Each Feature Set

	ACO	**Cepstrum**	**CepstrumBoW**
LR	0.524163	0.600902	0.538475
LDA	0.56369	0.583653	0.457783
KNN	0.397941	0.373775	0.454099
CART	0.472022	0.488099	0.465862
NB	0.475037	0.502309	0.459719
SVM	0.326139	0.326139	0.357944
ABC	0.529549	0.549869	0.534131
Kmc	0.112849	0.110909	0.149509

We tried different variations of the data set to identify accuracy changes in the trained models.

- Two actors act out in each session in the database and they are not involving in other sessions. Therefore, testing data will not include any voice that the algorithm has seen at the training session. Therefore, we mixed utterances among sessions and tried the same process with mixed data.
- Male and female voices have physical differences. Therefore, rather than developing a common model, we tried developing separate models for both male and female voices using separate data sets.
- IEMOCAP database has self-annotations for each utterance. Rather than using a composite annotation resulting from a set of annotations, we used self-annotation and trained models.

- Instead of using one cluster center for each emotion to create the codebook in Cepstral-BoW, we tried the same process with four clusters for each emotion. Therefore, there are 16 values per feature in the new codebook. By this, we tried to go deeper into data point spread and identifying sub ranges in different emotions.
- Instead of using *Happiness* and *Excitement* as a one category, we removed *Excitement* and retrained algorithms.

Re-training these classifiers with different changes in the data set in order to improve accuracy gave similar low accuracy results as well. Table 10 shows the maximum accuracy and the given classifier, each feature set got for every change we did. While other changes remain with similar or less accuracies, removing *Excitement* from *Happiness* and creating the dataset has gained an improvement. It significantly shows in Cepstrum-BoW feature set, improving the accuracy from 53% to 61%. It also has improved ACO and Cepstrum.

Table 10: Maximum Accuracy and the Given Classifier for Feature Sets in Multiclass Classification with Different Dataset Variations

	Using mixed sessions	**Male voice only**	**Female voice only**	**Self-Evaluation**	**Codebook with 4 centers**	**Without Excitement**
ACO	0.561429 (LDA)	0.509375 (LR)	0.561111 (ABC)	0.539137 (LR)	-	0.578328 (LDR)
Cepstrum	0.576429 (LR)	0.54875 (LR)	0.588889 (ABC)	0.572219 (LR)	-	0.616454 (LR)
Cepstrum BoW	0.531429 (LR)	0.531875 (ABC)	0.559722 (LR)	0.521946 (LR)	0.549095 (ABC)	0.61931 (ABC)

However, binary classification done on the same conditions to previously used feature sets gave significantly improved results. Table 11 shows the highest accuracy obtained from each binary classification from three different feature sets. Expecting *Happiness* and *Excitement* have features in common, classification was done considering them as a one Category. Ada Boost Classifier (ABC) gives the highest accuracy for identifying each emotion in every feature set. Further, the minimum among highest accuracy values is around 70%, which is to identify *Neutral* emotions in ACO category. This significant difference indicates that, binary classification works well than multiclass classification.

Table 11: Maximum Accuracy and the Classifier for Binary Classification with Three Feature Sets

Feature Set	**Hap+Exc**	**Sadness**	**Anger**	**Neutral**
ACO	0.761292(ABC)	0.841892(ABC)	0.858019(ABC)	0.707234(ABC)
Cepstrum	0.765535(ABC)	0.852258(ABC)	0.874523(ABC)	0.719502(ABC)
Cepstrum BoW	0.757348(ABC)	0.851056(ABC)	0.872919(ABC)	0.715932(ABC)
	Happiness	**Sadness**	**Anger**	**Neutral**
ACO	0.862028(SVM)	0.818885(ABC)	0.864280(ABC)	0.691331(ABC)
Cepstrum	0.862028(SVM)	0.821181(ABC)	0.884245(ABC)	0.712240(LR)
Cepstrum BoW	0.864828(ABC)	0.822759(ABC)	0.900690(ABC)	0.722069(ABC)

Further, after removing excitement category and taking the *Happiness* as a separate category, *Happiness* and *Anger* categories have taken significant improvements in their accuracy. *Excitement* has similarities with *Happiness* while having related features with *Anger* such as energy in the expression. This may have caused a gray area when distinguishing happy and anger. Specially, Cepstrum-BoW is created in a way that concerning the value range of emotion categories. Therefore, taking *Happy* as a separate category has stopped overlapping *Happy* and *Anger* ranges and improved the happy accuracy by 10%. Based upon these results, we continued using Happiness as a separate category.

3.5.3 Fusion

Fusion of multimodal results may improve the accuracy. In order to check the usability of fusion in the multiclass classifier, we fused above results. Expecting to improve the final result, we add a lexical component as well. In order to do that, we use sentiment analysis. Sentiment analysis of text is a method of identifying polarity of the speech. VADER (Valence Aware Dictionary and sEntiment Reasoner) [47] is an open source tool for sentiment analysis. It is specifically designed for sentiment analysis of the content social media. It provides positive, negative and neutral components of the sentiment of text. We used Vadar, which is a trained model, to analyze the text of the utterance to identify emotion. Therefore, the output of Vadar is directly used in the fusion.

Therefore, the multiclass, multimodal system consists of ACO classifier, Cepstrum classifier, Cepstral-BoW classifier and Vadar tool. First three models classify the audio, predicting an emotion category out of four emotions. In order to fuse these four outputs to a single emotion, we used a machine-learning approach. In the machine learning approach, we developed a fusion classifier, which takes outputs of emotion classifier as inputs. To train the fusion classifier in supervise-learning, we used session 4 data from the database to remove overfitting of data.

The maximum accuracy of fusion algorithm is 0.545319 given by SVM. This value became lower than Cepstrum and ACO individual accuracy values.

We used results of binary classification in the previous section to check behavior of fusion. In binary classification models, each feature set classified four times because of four emotion categories. Since we have three feature sets (ACO, Cepstrum and Cepstral-BoW) we got 12 outputs for a single utterance classification. In binary classification, we did not use lexical semantic analysis. Here we used a simple two tier rule based approach for fusion.

The first tier fuses same emotion binary classifications of three feature sets. For example, all three results from *Happiness* binary classification is fused. Therefore, first tier gives four values, each from one emotion category. We fused those results at the second tier. As shown in the Table 12, at the first tier, we decided the output based on the occurrence. Table 13 shows the combinations and relevant output that we used for the second tier fusion. For both tables, 'E' denotes an emotion while 'NE' is used to say no emotion detected. We tried adjusting the fusion rules in tables, however here we have shown the rule table, which gives the maximum fusion accuracy.

Table 12: Tier 1 Fusion Rules

Tier 1			
ACM	Cepstrum	Cepstral-BoW	Output 1
E	E	E	E
E	E	NE	E
E	NE	NE	E
NE	NE	NE	NE

Table 13: Tier 2 Fusion Rules

Tier 2				
Happiness(H)	Sadness(S)	Anger(A)	Neutral(N)	Output 2
0	0	0	0	N
0	0	0	1	N
0	0	1	0	A
0	0	1	1	A
0	1	0	0	S
0	1	0	1	S
0	1	1	0	A
0	1	1	1	A
1	0	0	0	H
1	0	0	1	H
1	0	1	0	H
1	0	1	1	H
1	1	0	0	e
1	1	0	1	e
1	1	1	0	e
1	1	1	1	e

In the Table 13, based on appearance of each emotions, resulting emotion is decided. 'e' condition denotes error scenarios which are hard to classify. Table 14 shows a sample result set and its fusion for a single utterance.

Table 14: A Sample Result Set in Binary Classification and Fusion

	Happiness	**Sadness**	**Anger**	**Neutral**
ACO	0	0	1	1
Cepstrum	0	0	1	0
Cepstrum-BoW	0	0	1	1
Tier 1 fusion	0	0	1	1
Tier 2 fusion		Anger		

We predicted emotions using trained binary classifiers performing on session 4 and 5 data set of IEMOCAP database. Then used that dataset for fusion. This time we did not include lexical values at this fusion due to its less performance obtained in previous fusion. Once tier 1 and 2 fusion is done, we got 990 correct predictions out of 2015 samples. Therefore the accuracy is 49%, which is very low compared to all the feature set level binary classifications.

3.6 Conversational Emotion Flow

Developed machine learning algorithm consists of identifying emotions (*Happiness, Sadness, Anger*, and *Neutral*) of utterances in conversations. For each utterance, we did binary classification for four emotions. Therefore, each utterance have four type of emotion measures. Depending on the utterance, we could identify the appearance of single or combination of emotions. We can use these results to visualize the emotion flow of conversations. Figures 4 & 5 shows the detected emotions of utterances in a dyadic conversation of speaker 1 and 2. Trend lines shows that how the each emotion is developing in the conversation. Figures 4 and 5 show that, two speakers are engaged in an angry conversation. Therefore, other emotions have slight or no appearance in the conversation. Figure 6 shows a comparison of *Anger* emotion flows of two speakers. This show that expressions of two speakers are involving in each other's emotion. Analyzing such patterns will open passages to identify abnormal behaviors of social robots. For example, social robots that are designed to provide emotional support such as accompany autism patients to minimize their hyper-activities and feelings should not participate in emotion exchange as shown in Figure 6.

Figure 4: Emotion Flow of Speaker 1 in a Dyadic Conversation

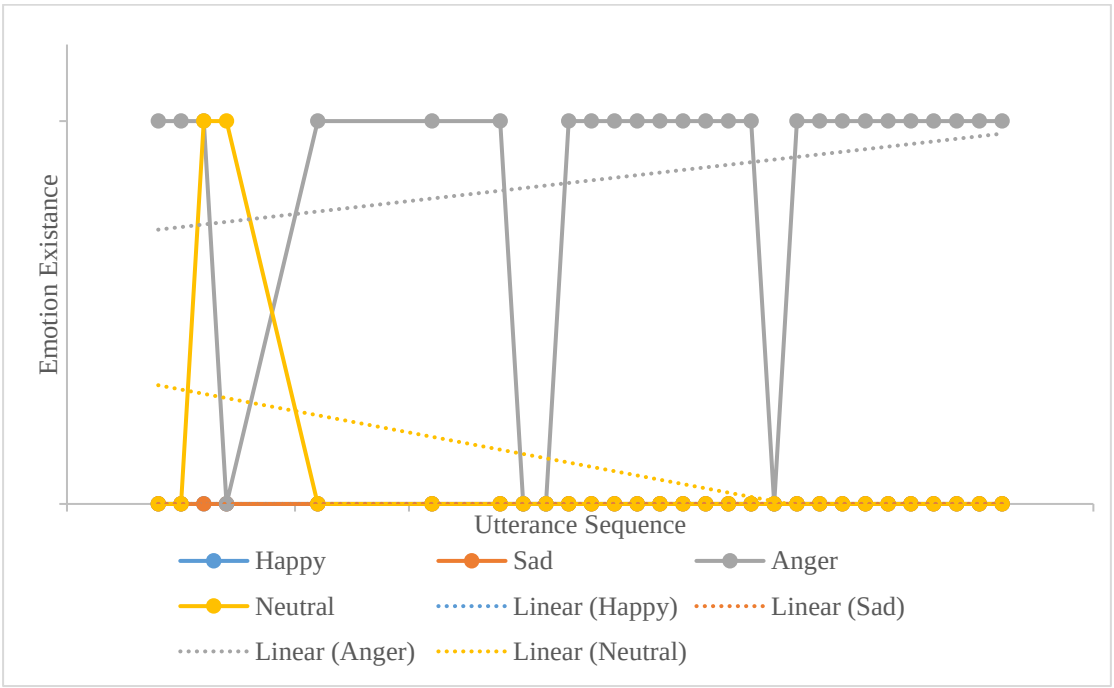


Figure 5: Emotion Flow of Speaker 2 in a Dyadic Conversation

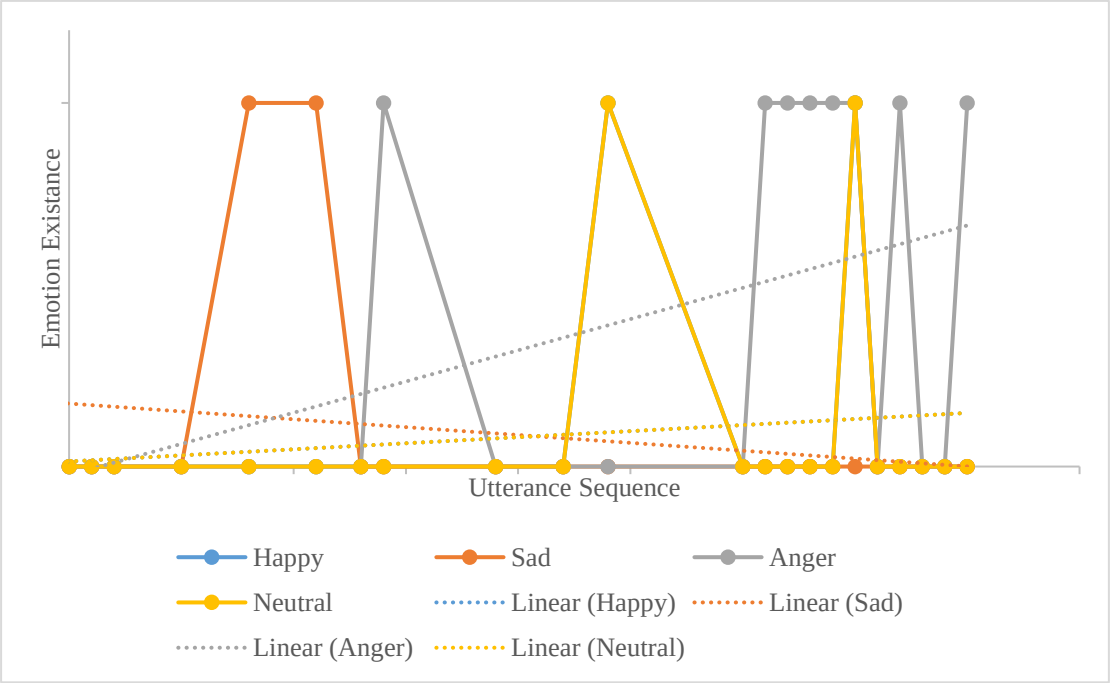
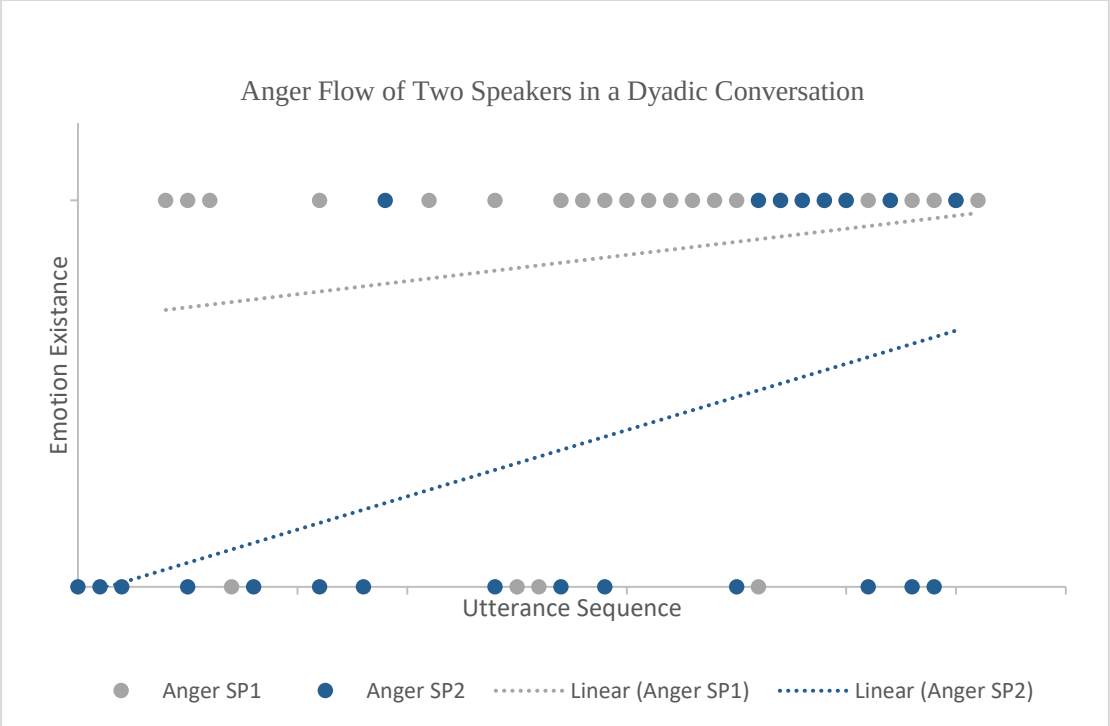


Figure 6: Anger Development in a Dyadic Conversation



4 CONCLUSION AND FUTURE WORK

In this research, we surveyed existing attacks and their effect on social robots. In addition, we researched vulnerabilities in systems that share similar characteristics with social robots. By analyzing both surveys, we could identify several attacks that create serious damage to social robots with existing or mutated versions of them. Among those attacks, we focused on abusing human emotions via human-social robot interaction. We are trying to identify abnormal social robot behavior via emotion detection. As a first step, we identified emotions of conversations in this research, which will be the building block of identifying abnormal social robot behavior via emotion detection.

In the emotion detection process, we have introduced a new method for feature filtering. This method plays competitively fast and accurate compared to other methods commonly used. Our method filters out comparatively a large number of features and leaves most effective features over. In the social robot setting, a real time emotion detection is needed in order to identify abnormalities in the behavior. Having less number of features will get less time for feature extraction from the audio, and use them in machine learning algorithm for prediction. However, in the proposed feature filtering method, the number of selected features slightly varies if filtered again with the same data set. As future work, we propose to look for improvements on finding cluster centers that can optimize filter accuracy for each emotion. In order to reduce the error in the cluster centers, we can repeatedly cluster a data set to obtain a set of cluster centers and use a second level clustering using that newly generated cluster centers dataset. Another way is, we can also use statistical methods such as mean or average. Further, a different clustering method can also be used for better accuracy.

Another proposed future work is to develop standard feature list for emotion detection by using proposed method. This proposed method has two dependencies.

- Data variance in the training data set
- Accuracy of annotation of the data set

A data set may not cover entire range of variance of the data for a certain emotion. If such new data values are found at the prediction, results will be erroneous. Therefore, a value set that has a full range is important. On the other hand, emotion categories that have close characteristics should be clearly identified in the annotation process. For example, Happy and Excitement have close values. To reduce these dependencies, we can combine different databases in order to incorporate much data

variance and normalize the annotation accuracy. Finally, we can develop a standard feature set to use in emotion detection.

In the multiclass and binary classification fusion processes, we obtained low accuracies. This is due to accumulating errors to the result, especially at the rule based fusion method. Therefore, we continued our work without fusion. We developed a system to identify a single set of features that gives higher accuracy in classification rather than having several such as ACO, Cepstrum, Cepstrum_BoW, and Lexical. On the other hand, it will not filter out information gathered for future levels. For example, we will have emotion flow for each emotion separately.

The interaction of humans with social robots can significantly impact the emotional state of a person. As a human companion, the SR is supposed to support humans. However, it can misbehave due to misconfiguration or because of an attack. We have developed a model, which identifies emotions in human-robot interactions. As the next stage, we propose to analyze emotion flow of conversations to identify both short-term conversational based harmful conditions as well as suspicious emotion patterns in the long-term basis.

5 BIBLIOGRAPHY

1. Sheng, S., et al., *Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions*, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2010, ACM: Atlanta, Georgia, USA. p. 373-382.

2. Kim, W., et al., *On botnets*, in *Proceedings of the 12th International Conference on Information Integration and Web-based Applications & Services*. 2010, ACM: Paris, France. p. 5-10.

3. Carrow, E.L., *Puppetnets and botnets: information technology vulnerability exploits that threaten basic internet use*, in *Proceedings of the 4th annual conference on Information security curriculum development*. 2007, ACM: Kennesaw, Georgia. p. 1-7.

4. Musavi, S.A. and M. Kharrazi, *Back to Static Analysis for Kernel-Level Rootkit Detection.* IEEE Transactions on Information Forensics and Security, 2014. **9**(9): p. 1465-1476.

5. Romana, S., et al. *Evaluation of open source anti-rootkit tools*. in *2013 Workshop on Anti malware Testing Research*. 2013.

6. Cui, W., et al., *Tracking rootkit footprints with a practical memory analysis system*, in *Proceedings of the 21st USENIX conference on Security symposium*. 2012, USENIX Association: Bellevue, WA. p. 42-42.

7. Xiongwei, X. and W. Weichao. *Rootkit detection on virtual machines through deep information extraction at hypervisor-level*. in *2013 IEEE Conference on Communications and Network Security (CNS)*. 2013.

8. Yin, H., et al., *Panorama: capturing system-wide information flow for malware detection and analysis*, in *Proceedings of the 14th ACM conference on Computer and communications security*. 2007, ACM: Alexandria, Virginia, USA. p. 116-127.

9. Yin, H., Z. Liang, and D. Song, *HookFinder: Identifying and Understanding Malware Hooking Behaviors*. 2008.

10. Denning, T., et al., *A spotlight on security and privacy risks with future household robots: attacks and lessons*, in *Proceedings of the 11th international conference on Ubiquitous computing*. 2009, ACM: Orlando, Florida, USA. p. 105-114.

11. Jeong, S.-Y., et al., *A Study on ROS Vulnerabilities and Countermeasure*, in *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 2017, ACM: Vienna, Austria. p. 147-148.

12. Templeton, S.J., *Security aspects of cyber-physical device safety in assistive environments*, in *Proceedings of the 4th International Conference on PErvasive Technologies Related to Assistive Environments*. 2011, ACM: Heraklion, Crete, Greece. p. 1-8.

13. Lin, H., et al., *Safety-critical cyber-physical attacks: analysis, detection, and mitigation*, in *Proceedings of the Symposium and Bootcamp on the Science of Security*. 2016, ACM: Pittsburgh, Pennsylvania. p. 82-89.

14. Junejo, K.N. and J. Goh, *Behaviour-Based Attack Detection and Classification in Cyber Physical Systems Using Machine Learning*, in *Proceedings of the 2nd ACM International Workshop on Cyber-Physical System Security*. 2016, ACM: Xi'an, China. p. 34-43.

15. Mitchell, R. and I.-R. Chen, *A survey of intrusion detection techniques for cyber-physical systems.* ACM Comput. Surv., 2014. **46**(4): p. 1-29.

16. Al-Sarawi, S., et al. *Internet of Things (IoT) communication protocols: Review*. in *2017 8th International Conference on Information Technology (ICIT)*. 2017.

17. Mustapha, H. and A.M. Alghamdi, *DDoS attacks on the internet of things and their prevention methods*, in *Proceedings of the 2nd International Conference on Future Networks and Distributed Systems*. 2018, ACM: Amman, Jordan. p. 1-5.

18. Bertino, E. and N. Islam, *Botnets and Internet of Things Security.* Computer, 2017. **50**(2): p. 76-79.

19. Park, J., et al., *Study of Car Dash Cam Security Vulnerabilities*, in *Proceedings of the 4th International Conference on Information and Network Security*. 2016, ACM: Kuala Lumpur, Malaysia. p. 73-76.

20. Miller, J., A.B. Williams, and D. Perouli, *A Case Study on the Cybersecurity of Social Robots*, in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 2018, ACM: Chicago, IL, USA. p. 195-196.

21. Giaretta, A., M.D. Donno, and N. Dragoni, *Adding Salt to Pepper: A Structured Security Assessment over a Humanoid Robot*, in *Proceedings of the 13th International Conference on Availability, Reliability and Security*. 2018, ACM: Hamburg, Germany. p. 1-8.

22. Bj, et al., *Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge.* Speech Commun., 2011. **53**(9-10): p. 1062-1087.

23. Roh, Y.-W., et al., *Novel acoustic features for speech emotion recognition*. Vol. 52. 2009. 1838-1848.

24. Aldeneh, Z., et al., *Pooling acoustic and lexical features for the prediction of valence*, in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 2017, ACM: Glasgow, UK. p. 68-72.

25. Rong, J., et al. *Acoustic Features Extraction for Emotion Recognition*. in *6th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007)*. 2007.

26. Savran, A., et al., *Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering*, in *Proceedings of the 14th ACM international conference on Multimodal interaction*. 2012, ACM: Santa Monica, California, USA. p. 485-492.

27. Tian, L., J.D. Moore, and C. Lai, *Recognizing emotions in spoken dialogue with acoustic and lexical cues*, in *Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents*. 2017, ACM: Glasgow, UK. p. 45-46.

28. Chuang, Z.-J. and C.-H. Wu, *Multi-Modal Emotion Recognition from Speech and Text*. Vol. 9. 2004.

29. Anagnostopoulos, C.-N. and T. Iliou, *Towards Emotion Recognition from Speech: Definition, Problems and the Materials of Research*. 2010. p. 127-143.

30. Weenink, P.B.D., *Praat: doing phonetics by computer [Computer program]*. 2018.

31. Chen, S., et al. *Speech emotion classification using acoustic features*. in *The 9th International Symposium on Chinese Spoken Language Processing*. 2014.

32. Rozgic, V., et al., *Emotion Recognition using Acoustic and Lexical Features*. Vol. 1. 2012.

33. M.F. Porter (Computer Laboratory, C., UK), *An algorithm for suffix stripping.* Program, 2006. **40**(3): p. 211-218.

34. Jin, Q., et al. *Speech emotion recognition with acoustic and lexical features*. in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015.

35. Li, J., et al., *Feature Selection: A Data Perspective.* ACM Comput. Surv., 2017. **50**(6): p. 1-45.

36. Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python.* J. Mach. Learn. Res., 2011. **12**: p. 2825-2830.

37. Hozjan, V. and Z. Kacic, *Context-Independent Multilingual Emotion Recognition from Speech Signals*. Vol. 6. 2003. 311-320.

38. Yu, L. and H. Liu, *Feature selection for high-dimensional data: a fast correlation-based filter solution*, in *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*. 2003, AAAI Press: Washington, DC, USA. p. 856-863.

39. Cheung, Y. and H. Jia. *Unsupervised Feature Selection with Feature Clustering*. in *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. 2012.

40. Chormunge, S. and S. Jena, *Correlation based feature selection with clustering for high dimensional data.* Journal of Electrical Systems and Information Technology, 2018. **5**(3): p. 542-549.

41. Bj, et al., *AVEC 2012: the continuous audio/visual emotion challenge - an introduction*, in *Proceedings of the 14th ACM international conference on Multimodal interaction*. 2012, ACM: Santa Monica, California, USA. p. 361-362.

42. Livingstone, S.R. and F.A. Russo, *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English.* PLOS ONE, 2018. **13**(5): p. e0196391.

43. Planet, S. and I. Iriondo. *Comparison between decision-level and feature-level fusion of acoustic and linguistic features for spontaneous emotion recognition*. in *7th Iberian Conference on Information Systems and Technologies (CISTI 2012)*. 2012.

44. Busso, C., et al., *IEMOCAP: Interactive emotional dyadic motion capture database.* Journal of Language Resources and Evaluation, 2008. **42**: p. 335-359.

45. Eyben, F., et al., *Recent developments in openSMILE, the munich open-source multimedia feature extractor*, in *Proceedings of the 21st ACM international conference on Multimedia*. 2013, ACM: Barcelona, Spain. p. 835-838.

46. Schuller, B., et al., *The INTERSPEECH 2010 paralinguistic challenge*. 2010. 2794-2797.

47. Hutto, C.J. and E.E. Gilbert, *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.* Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI.

6 APPENDIX

6.1 Survey 1: Attacks Methods and Social Robot

Full version of the table, Current Threats and Protections for Social Robots in the Survey 1 attached is below.

Table 15: Current Threats and Protections for Social Robots - Full Version

		Behavior	Current Protection Tools and Techniques	Possibility of existing techniques	Limitations to use existing tools	Possible Relations to Social Robots (SR)
Web Phishing		Acquire personal information using, Social Engineering (fake links in emails, fake web pages, fake links in web advertisements), malicious redirections, XSS attacks, Cloning	2 factor authentication social engineering prevention methods Anti-Phishing toolbars which compare visiting sites with phishing sites Anti-Spam email solutions (ZoEmail, Vanquish Anti-Spam, Vanquish Anti-Spam, Spamfence, Spam Arrest, AlienCamel, 0Spam)	Two-factor authentication may be used if the robot have access to the secondary device. Separately hosted anti-spam email solutions connected to the email account may support in general.	Incompatibility of preventions techniques for social engineering like making web pages personally recognizable, user awareness to use human instinct. Browser based solutions are not supporting since SR will not have browsers.	Most of the phishing attacks are based on social engineering. In SR setting, some of them are obsolete, and some are beyond the control of SR. Ex: robots can identify visually similar but fake URLs while robots will not identify fake emails. Further, if attacker could route SR to a fake web page by other mean like setting up a temporary DNS, SR cannot identify fake visuals on that web site.
Denial of Service	**Other**	Overwhelming system resources like Network bandwidth, Server memory, Application exception handling mechanism, CPU usage, Hard disk space, Database space and Database connection pool. This will lead to disturb services.	Traffic management, load balancing tools, Response Rate Limiting(RRL), Collection of reverse proxies, SYN cookies, Null-Routing by ISP	Null routing by ISP can be used but it is not in the controllable zone	Limited resources and the nature of SR prevent using bulky systems such as, Load balancers, traffic management systems, and reverse proxies. Network device configurations like RRL cannot be expected from users	This can shutdown services, which provide essential services or entire device.
	SYN flood	Sending SYN with not reachable ip's to fill backlog queue, which make no room for new connections	Decreasing TCP connection timeout in devices, MS Windows mechanism which monitors number of half opened connections and number of refused connections Check point firewall	System configurations like TCP timeout can be used.	Firewall protection is incompatible	This can isolates the robot, so that any IoT device or anybody cannot connect to the robot, especially from outside when owner monitoring.

	UDP Flood	Flooding with UDP to perform a Dos attack	Firewall configuration to filter malicious UDP, Load balancing, Configuring routers(thresholds) to filter UDP flooding	Firewalls and load balancers are incompatible	This can overwhelm ports making unresponsive to clients and shutdown services. Further, this may cause performance interruption to other essential services like security.
	Smurf Flood Attack	Broadcasting many ICMP request with spoofed source address. Victim receive flooding.	Shutting Broadcast addressing features in routers and firewalls, configuring end nodes not to respond ICMP requests	Node configuration is hard for SR users Firewalls will not available for SR setting	SR may provide set of essential services via internet or in-house network. Ex: Online home monitoring, in-house device control. Attacker who is connected to the internet or connected to the home network can use this attack to shutdown such services.
Malicious software (Malware)	**Virus, Trojan horse, Spyware, Ransomware, Backdoors/remote access, Trojan, Bugs**	Virus - execute itself and spread by infecting other programs and files Trojan horse - Pretending as a legitimate program and get into the system Spyware - Collect data n info without permission Ransomware - Infect and encrypt data Backdoors/remote access Trojan - Secretly creates backdoors to connect remotely Bugs - making processes to giving undesired outcomes	Anti-social Engineering based identification and prevention methods Anti-virus software firewalls Many malware detectors	The systems like firewalls does not compatible with SR, Anti-virus and malware detectors needs to be light weighted if inserting to SR. Still that will need many resources and that affect the system performance. Further defending methods for social engineering attacks can also become unsupportive.	SR environment is vulnerable for these malicious activities and there can be many ways malicious codes getting into the SR system. They can make more harmful actions based on the nature of SR. For example spyware can collect more personal information, rootkits can manipulate movements of SR

Internet Worms	A malicious program that can replicate it by itself and can travel across the network without human action. Apart from that, the basic malicious function can be changed based on the purpose it created.	User based identification and prevention methods, Anti-virus software, firewalls, Timely system updates	Automatic timely system updates will closedown system loop-holes as much as possible	Prevention methods based on user will not help. Bulky anti-virus software will not support SR setting. Firewalls will not be available in SR setting	Worms can spread in SR setting and deliver malicious activities. They can make system issues, utilize limited resources, information theft in the robotic system
Botnet Attacks	become a spam bot that render advertisement on websites become web spider that scrapes server data distributing malware disguised as popular search items on download sites Participating in Ddos attacks Spamming Sniffing traffic Key logging Participating in manipulating online polls/games etc.	Honeypot/Honeynet, IRC tracking, DNS tracking, DNS tracking, Firewall protection, IDS/IPS		Hard to implement prevention tools inside SR since they are light weighted and mobile. In addition, SR are not working in heavily secured sophisticated networks. Therefore, Honeypot kind of solutions are also helpless.	SR can become a zombie robot. Its setup support botnet activities. There can be many idle robots to utilize and on the other hand, robot performance can be degraded because of botnet activities. These activities may target outside targets or they can harm the user as well. There can be a little chance to become a target of a botnet.

	Rootkit Attacks	Intruder get root access to the system and he can do anything that an administrative user (root) can do.	Anti-social engineering concepts for users like not clicking on unknown emails, attachments, links, installing certified software, etc. User based or Kernel based anti-rootkit software (Linux: chkrootkit, Lynis, ISPProtect, rkhunter; Android: Lookout, malwarebytes, rootkit detector; Windows: Milano, Webroot, F-Secure blacklight; Windows RT: Webroot; IOS: Lookout, Webroot; OS X: Osquery, Webroot)		Most of the social robots do not have performance capability as other computers since their resources are limited. Therefore running a rootkit scanner may cause additional hit to its performance. In addition, most of the rootkit detectors are offline, that is not a proper solution since SR need real time solutions, and users may not need to have such offline solutions. Hence, this may not be a perfect solution. Limitations in prevention methods to social engineering attacks. On the other hand bringing them in to a separate central system like a cloud for monitoring may create complex infrastructure requirements with additional heavy processes.	An intruder getting root access will allow himself to do number of harmful actions in SR. Different robots have different capabilities. Therefore based on the capabilities intruder may get more chances. Higher capabilities higher the risk. Ex: A robot who can move around the house and pick door locks can let the intruders physically come in. A robot who has cameras, microphones may let intruders to spy.
	Mobile Malware	Malicious codes get inside from different ways, and perform malicious activities on mobile devices	Installing apps only from trusted sources, using mobile protections systems, prevent jailbreaking, Timely system updates,	Most of the existing tools and techniques can be used in SR to secure them		There are several operating systems used in SR. If SR start using mobile platforms, mobile malware can affect SR. However, SR are having different design than other mobile devices. Therefore some mobile malware can be more harmful than expected

Zero-day Attack	Exploit previously unknown vulnerabilities	Updating/Patching	Automatic updating and Patching can be done	Knowledge of users of SR to do updates and patching	SR can have this type of attacks as other systems. In addition, in the similar way, designers have to foresee such vulnerabilities before attackers, and take necessary actions.
Session Hijacking	Through session fixation, sidejacking, cross-site scripting or physical access become a middle man	Tunneling, IPSec, Encryption(SSl/TLS), Using long random numbers or strings, Regenerating the session id after login, changing cookie values with each and every request, secondary checks like ip matching	Protocol security and application level secure options can be used		This can happen in SR setting in the similar way that happens in other contexts. That will leak information from SR and may be used to spy
OS Vulnerabilities	A malware or network based attack through OS loopholes by Stealing information, destructive operations using codes, scripts, active content and etc.	Regular software patching	Automatic software patching	Cannot expect user to install new patches timely, because the intended users may range from small children to old patients	OS vulnerability attacks are acting in a similar way for SR as well. But since SR of should support more hardware (Motors, sensors) than a computer, it could have more vulnerability chances than a general computer
Platform Vulnerabilities	Steal information, destructive operations using codes, scripts, active content and etc.	System updates and patches	Frequent auto updates and patching will help to mitigate the risk of such vulnerabilities	Updates and patches need to done automatically since knowledge of user in that context cannot be expected in SR scenario	System vulnerabilities will be there for many systems including SR. That can leave spaces to use by attackers in different ways.
Application Vulnerabilities	Backdoors are let open by applications to attacker for executing commands and access data.	Regular software patching, using antivirus software	Regular updates and patching in API and factory developed applications	Unlike Apple, Google or Microsoft there is no structure for validating the security concerns of new apps developed using API's, before they get available for ROS or related OS in robots	Some robotic platforms provide APIs to develop user applications for the robot. There is no certificate for security of those apps. These new and existing apps may have a big chance to include many vulnerabilities

Buffer Overflow		This is a general error but an attacker can use it to exploit program buffer. Setting more data to the buffer than it is defined can leads to crash the program or set new values to adjacent variables. Stack-based overflow, Heap-based overflow, Integer overflow, String overflow, Unicode overflow	Updating/Patching, Safe coding, Compiler based detection(setting random values aka canaries to check), Operating system based non-executable stacks and ASLR(Address Space Layout Randomization)	Auto-updating and patching will remove loop-holes, Code standards		This could affect different programs running inside the robot; hence, different service may be shut down. Thus can be critical if the program is critical
Spoofing	**Packet Spoofing**	Forge the source and pretend as a trusted src. This can lead to session hijacking or intercepting network traffic.	Configure ACL(Access Control List) to Deny incoming packets if source address is allocated to your network, Deny outbound packets if source address is not allocated to your network Unicast Reverse Path Forwarding (uRPF); discarding IP packets that lack a verifiable IP source address in the IP routing table. IP Source Guard is a Layer 2 security	ACL and uRPF	Cannot expect network configurations from general users	This can be used to steal information from SR or can use this to spy on the user or environment.
	MAC Address Spoofing	Change MAC to some authorized MAC and pretend as someone else in the network. Some OS are allowing	Cisco port security		Network configurations cannot be expected from SR users	Regardless of the systems, this can happen. SR will send traffic to the intruder. He will read the information. This can lead to other attacks as well.

		changing the MAC. Related to Data Link Layer.				
	IP Address Spoofing Attack	Change IP to some authorized IP and pretend as someone else in the network. Related to Network Layer.	Configure ACL(Access Control List) to Deny incoming packets if source address is allocated to your network, Deny outbound packets if source address is not allocated to your network			
Peer-to-peer (P2P) File Sharing		Installing malicious codes, access through opening ports, stealing info, dos	Anti-virus software, firewalls		Protecting FTP by using firewalls and virus guards is difficult in SR	SR may have FTP connections with user smart phone or personal computer. Other than that based on different application requirements FTP connections may be needed. There is a possible risk of malicious activities on peer to peer connections
Scanning and probing		Available services and ports will be revealed, This is a legitimate audit function as well.	There are different commercial tools that are used to probing in good purpose and bad purpose. Using them in protecting way and identify network vulnerabilities. Keeping close the port that are not using. Using firewalls and IDS/IPS to identify spoofing attacks	By default, keeping non-using ports closed	probing to check vulnerabilities and securing ports for security purposes is better for large service providers, but it does not supporting for SR. Users will not perform such administrative tasks for SR.	SR services may uses number of ports for its services. Based on applications SR has ports may get unsecured. Insecure, kept open ports will expose the robot to the outside, which can lead to an attack
Traffic (Packet) Sniffing		Reading packets in the middle of the network. This is also a legitimate function as well.	There are different tools that can capture the traffic and convert data into human readable format.	Encryptions and Light weighted anti-sniffing tool can be used	though a sniffer is identified, there may be less chances to take actions	Regardless of the system type, this attack can happen. Any data traveling through the network is at risk at this point. In SR

		Using them to identify vulnerabilities and take precautions. Encrypting (SSL, TSL) data, but this still reveals src and dst details Using commercial ant sniff software		in a home network with SR users	setting it home network consists of more data than a regular home network.
Eavesdropping	Unauthorized real time interception of a private communication. Wiretapping is the method for conventional telephones, for VoIP, there are sniffing tools to read the IP packets.	Encrypting data before transmit., avoiding public networks and using VPN	Encryption can be used in SR to make the communication secure		SR may involve in communication like phone calls, messages. Therefore VoIP (including conventional telephony traffic in a way, since SR may be capable of connecting phones. Ex: vehicle audio support for calls) and other messaging traffic will transfer through SR to user or out from user. Attackers can eavesdrop user's communication via SR
ARP Poisoning	Replace the MAC to an existing IP is poisoning. This can be used to act as a man in middle, setting same MAC to several ip's to create a DOS attack, steal session id and hijack the session. ARP spoofing software include ARPspoof, Cain & Abel, ARPpoison and Ettercap.	Packet filtering- inspect packets as they are transmitted across a network. ARP spoofing detection software- Programs inspecting and certifying data before it is transmitted and blocking data that appears to be spoofed. Cryptographic network protocols- Transport Layer Security (TLS), Secure Shell (SSH), HTTP Secure (HTTPS) and other secure communications protocols bolster ARP spoofing	Using cryptographic protocols is a possible solution related to SR	Hard to expect network monitoring in a home network	Home networks where SR are residing are not most secured with firewalls and other protection mechanisms. There is a more chance to be attacked in such network. Resulting MITM, DoS, Session hijacking, Spoofing, Sniffing and many other attacks on SR

		attack prevention by encrypting data prior to transmission and authenticating data when it is received.			
Broadcast Storm	Overwhelming by continuous multicast or broadcast	Judicious use of firewalls, Better network configurations, storm controls		Firewalls and network configuration to prevent such attacks cannot be expected in a home network	Attacker can shut down the robots connectivity and that can be a security threat, and will interrupt services
Offline Authentication attacks(Brute Force, Dictionary, Rainbow table)	Use stolen or other available information to guess passwords and PINs not trying with online systems	Using CAPTCHA's, account locking, progressive delays, salting, Strong password structures, biometrics, tokens	Almost all of the tools and techniques are supporting to SR setting	Passwords are mostly stored	SR can be a single access point for many accounts of the user. Once the authority is given by user, SR will access these accounts frequently with or without the telling the user. This attack will work in SR similar to other systems.
DHCP Attacks	With fake MAC, exhaust dhcp srv and setup rough dhcp (DHCP Starvation attack). Attacker can act like a trusted node like DNS or default gateway (DHCP spoofing attack)	Port security		Network configurations cannot be expected from SR users	Regardless of the systems, this can happen. SR will send traffic to the intruder. He will read the information. This can lead to other attacks as well.
MAC Table Overflow	Make the switch flooding using fake mac address. Router flushes the mac table and starts broadcasting all the packets. Attacker can receive all the frames.	Cisco port security		Heavy network security and administration are not available in home network setting,	Regardless of the device connected to the network, this problem happens. Attacker can get data travelling around specially communication between SR and IoT devices, User mobile phone, Laptop or computer. That makes this attack in SR setting more harmful
Weak Passwords	Breaking simple passwords and get access	Strong password structures	Strong password structures		Attacks for weak passwords can happen in SR as they use password-protected accounts. Users of SR can be vary from

					child to elderly person. The chances to have a weak password is high
Unapproved apps and portable devices	Malicious codes get inside through untrusted app installation or physically connecting devices like USB storages	Anti-Virus guards, Controlling restrictions of access	Access controlling, Providing secure and trusted platform to install apps	Resource limitations to use inbuilt virus guard or detection systems	Some robotic platforms provide APIs to develop user applications for the robot. There is no certificate for security of those apps. Vulnerabilities in these apps may invite more attacks. Further, attackers can also use these APIs to bait SR users. Devices connecting to SR may be a media to get into the system as well
Data loss from lost or stolen device	By their nature, handy portable devices can be physically stolen	Locating mechanisms using GPS to locate and password/fingerprint/face recognition protection till it found, Online data erasing mechanisms	Current protection mechanism are still valid for SR setting. To make it more secure, encrypting can be used for storing data in SR		Often SR authenticate the user using face recognition. If such security methods are accurate enough, that could prevent outsiders or people who stole it using the SR. That will prevent data access somewhat. In addition, unlike mobile phones or laptops, SR does not provide much easier file system accessing mechanism. Hence it will make it bit complex to access data from a stolen SR. Attacker may need to get hardware level to access data or reset configurations.

6.2 Binary Classification Data

Additional data for binary classification with *Happiness* and *Excitement* for ACO, Cepstrum and Cepstrum-BOW feature sets is added here.

Table 16: Binary Classification for ACO with Happiness and Excitement

	ACO			
	Hap+Exc	Sadness	Anger	Neutral
LR	0.743283	0.830385	0.842657	0.684188
LDA	0.755187	0.833837	0.848411	0.690354
KNN	0.714122	0.785886	0.789704	0.641992
CART	0.68725	0.79047	0.788192	0.639285
NB	0.624686	0.651575	0.72487	0.575197
SVM	0.72794	0.787395	0.810805	0.673861
ABC	0.761292	0.841892	0.858019	0.707234
Kmc	0.178718	0.11863	0.105097	0.10893

Table 17: Binary Classification for Cepstrum with Happiness and Excitement

	CEP			
	Hap+Exc	Sadness	Anger	Neutral
LR	0.758631	0.838441	0.866453	0.724861
LDA	0.738706	0.825385	0.871055	0.69111
KNN	0.711054	0.750576	0.825389	0.594419
CART	0.704556	0.782038	0.819633	0.668821
NB	0.610516	0.775507	0.709892	0.5821
SVM	0.72794	0.787395	0.810805	0.673861
ABC	0.765535	0.852258	0.874523	0.719502
Kmc	0.178473	0.097496	0.077121	0.074829

Table 18: Binary Classification for Cepstrum-BoW with Happiness and Excitement

	CEP_BOW			
	Hap+Exc	Sadness	Anger	Neutral
LR	0.702146	0.814833	0.839565	0.675131
LDA	0.654405	0.746406	0.771141	0.632539
KNN	0.698106	0.80795	0.83724	0.669354
CART	0.686031	0.79009	0.833792	0.663584
NB	0.588267	0.76367	0.671656	0.571015
SVM	0.722254	0.776872	0.826899	0.683177
ABC	0.757348	0.851056	0.872919	0.715932
Kmc	0.109265	0.110378	0.094881	0.09315

Additional data for binary classification without *Excitement* for ACO, Cepstrum and Cepstrum-BOW feature sets is added here.

Table 19: Binary Classification for ACO without Excitement

	ACO			
	Happiness	Sadness	Anger	Neutral
LR:	0.846137	0.812536	0.844286	0.659566
LDA:	0.827972	0.808453	0.848813	0.671802
KNN:	0.844778	0.739428	0.779817	0.59782
CART:	0.779385	0.751228	0.794774	0.632351
NB:	0.396687	0.648182	0.77986	0.562022
SVM:	0.862028	0.74355	0.779831	0.614591
ABC:	0.852501	0.818885	0.86428	0.691331
Kmc:	0.1411	0.123019	0.164809	0.116154

Table 20: Binary Classification for Cepstrum without Excitement

	CEP			
	Happiness	Sadness	Anger	Neutral
LR:	0.83023	0.824342	0.872908	0.71224
LDA:	0.801213	0.797112	0.867892	0.672768
KNN:	0.852954	0.699981	0.80754	0.567919
CART:	0.775309	0.776189	0.827061	0.659111
NB:	0.565564	0.746242	0.748513	0.601039
SVM:	0.862028	0.74355	0.779831	0.614591
ABC:	0.853836	0.821181	0.884245	0.696345
Kmc:	0.103965	0.144443	0.144938	0.100765

Table 21: Binary Classification for Cepstrum-BoW without Excitement

	CEP_BOW			
	Happiness	Sadness	Anger	Neutral
LR:	0.776552	0.784828	0.837931	0.669655
LDA:	0.689655	0.677931	0.791034	0.613793
KNN:	0.846207	0.781379	0.828966	0.629655
CART:	0.770345	0.764828	0.848966	0.658621
NB:	0.650345	0.755172	0.718621	0.657931
SVM:	0.863448	0.734483	0.792414	0.615862
ABC:	0.864828	0.822759	0.90069	0.722069
Kmc:	0.109655	0.154483	0.175172	0.111034

6.3 Fusion Data

Table 22: Variations of Tier 2 Fusion Rules

Tier 2 Fusion								
Hap(H)	Sad(S)	Ang(A)	Neu(N)	O/P v1	O/P v2	O/P v3	O/P v4	O/P v5
0	0	0	0	E	N	N	N	N
0	0	0	1	N	N	N	N	N
0	0	1	0	A	A	A	A	A
0	0	1	1	A	A	A	A	A
0	1	0	0	S	S	S	S	S
0	1	0	1	S	S	S	S	S
0	1	1	0	A	A	A	A	A
0	1	1	1	E	A	S	A	A
1	0	0	0	H	H	H	H	H
1	0	0	1	H	H	H	H	H
1	0	1	0	E	H	A	H	H
1	0	1	1	E	H	A	H	H
1	1	0	0	E	E	E	N	H
1	1	0	1	E	E	E	N	H
1	1	1	0	E	E	E	N	H
1	1	1	1	E	E	E	N	H

6.4 Feature Filtering Data

In order to identify best set of features, different combinations of *Inclusion* and *Min_Distance* is tested. Below table contains the results of the test.

Table 23: Binary Classifier Accuracy for Inclusion and Min_Distance Combinations - Full

		15_all	**30_all**	**40_2**	**40_2.5**	**40_3**	**50_2**	**65_2**	**80_2**	**80_2.5**	**80_3**	**80_3.5**	**80_4**
Anger	LR	0.778922	0.787995	0.781647	0.781645	0.781647	0.863367	0.871086	0.871086	0.891514	0.891518	0.876524	0.868809
	LDA	0.778922	0.776645	0.773015	0.77165	0.772096	0.852478	0.876991	0.876991	0.885613	0.883344	0.871989	0.867896
	KNN	0.738943	0.752616	0.753949	0.75123	0.744883	0.769383	0.765294	0.765294	0.771187	0.770294	0.798447	0.827042
	CART	0.675872	0.744436	0.711271	0.693579	0.695378	0.789375	0.818877	0.818877	0.811156	0.821615	0.832497	0.814369
	NB	0.775296	0.545607	0.668612	0.753501	0.761674	0.664531	0.752631	0.752631	0.762153	0.777585	0.789846	0.804823
	SVM	0.778013	0.776656	0.779831	0.779831	0.774381	0.779831	0.779831	0.779831	0.779831	0.779831	0.779831	0.779831
	ABC	0.770757	0.793459	0.770302	0.769848	0.774829	0.852009	0.863361	0.863361	0.874729	0.862923	0.868824	0.861104
	Kmc	0.004988	0.040864	0.113908	0.107972	0.155261	0.098005	0.142137	0.142137	0.089478	0.056789	0.204757	0.108371
		15_all	**30_all**	**40_2**	**40_2.5**	**40_3**	**50_2**	**65_2**	**80_2**	**80_2.5**	**80_3**	**80_3.5**	**80_4**
Happiness	LR	0.862028	0.862028	0.859755	0.86021	0.862028	0.860666	0.859303	0.862016	0.862026	0.860671	0.862028	0.862028
	LDA	0.861573	0.86021	0.859755	0.860662	0.862028	0.858392	0.856586	0.855683	0.857483	0.862493	0.862937	0.862028
	KNN	0.844317	0.851135	0.847954	0.848852	0.855222	0.854761	0.852046	0.850237	0.850216	0.841604	0.847511	0.847517
	CART	0.762145	0.757618	0.763046	0.752158	0.837964	0.751236	0.757598	0.788937	0.805278	0.788056	0.784864	0.751703
	NB	0.822536	0.314099	0.383573	0.799831	0.842065	0.411664	0.485202	0.574192	0.624585	0.657273	0.684957	0.803924
	SVM	0.859305	0.859303	0.862028	0.862028	0.85885	0.862028	0.862028	0.862028	0.862028	0.862028	0.863387	0.860212
	ABC	0.858396	0.857483	0.852046	0.857038	0.857493	0.845224	0.843861	0.844774	0.851131	0.852042	0.853865	0.857046
	Kmc	0.05448	0.073571	0.13261	0.105249	0.011783	0.077633	0.146028	0.102653	0.062563	0.148863	0.139334	0.100695

		15_all	30_all	40_2	40_2.5	40_3	50_2	65_2	80_2	80_2.5	80_3	80_3.5	80_4
Neutral	LR	0.613682	0.620037	0.612779	0.614591	0.614591	0.65094	0.699556	0.723137	0.715411	0.709044	0.697234	0.641364
	LDA	0.617314	0.612314	0.594169	0.615045	0.614591	0.66594	0.682302	0.696357	0.70225	0.716288	0.699039	0.657711
	KNN	0.58827	0.548797	0.560594	0.543815	0.551503	0.592842	0.602369	0.597371	0.583318	0.588737	0.655033	0.663638
	CART	0.563291	0.620506	0.589687	0.535142	0.521981	0.625072	0.623239	0.630967	0.637807	0.6532	0.660006	0.655033
	NB	0.503455	0.540226	0.502548	0.498451	0.4939	0.538433	0.545685	0.564724	0.572443	0.576989	0.567445	0.566522
	SVM	0.610936	0.600072	0.614591	0.614591	0.611868	0.614591	0.614591	0.614591	0.614591	0.614591	0.61414	0.674089
	ABC	0.625051	0.649095	0.636837	0.586886	0.595961	0.683626	0.681366	0.711752	0.710847	0.716759	0.698163	0.674101
	Kmc	0.012721	0.093544	0.1031	0.115732	0.066742	0.109749	0.113429	0.127158	0.122505	0.152534	0.122145	0.113011
		15_all	**30_all**	**40_2**	**40_2.5**	**40_3**	**50_2**	**65_2**	**80_2**	**80_2.5**	**80_3**	**80_3.5**	**80_4**
Sadness	LR	0.742641	0.784831	0.80935	0.81253	0.794803	0.826619	0.828885	0.82977	0.831139	0.830247	0.825257	0.7971
	LDA	0.74355	0.784377	0.808433	0.803453	0.787092	0.823877	0.826608	0.825226	0.833416	0.827966	0.821166	0.80391
	KNN	0.721312	0.719467	0.77849	0.791203	0.796205	0.753509	0.684066	0.708106	0.734926	0.724029	0.755804	0.744436
	CART	0.727186	0.768063	0.768955	0.781201	0.779401	0.781222	0.782589	0.779375	0.779864	0.760339	0.769426	0.769854
	NB	0.492489	0.573747	0.726281	0.714938	0.714029	0.727643	0.736705	0.765738	0.743517	0.736259	0.735346	0.738525
	SVM	0.734033	0.74673	0.792102	0.820701	0.814364	0.74355	0.74355	0.74355	0.74355	0.74355	0.74491	0.773947
	ABC	0.775755	0.803023	0.811174	0.81072	0.806625	0.818902	0.825241	0.827974	0.826608	0.826162	0.808457	0.788947
	Kmc	0.131098	0.091292	0.151074	0.13252	0.121168	0.093023	0.070825	0.109698	0.123593	0.137577	0.189299	0.138453

Table 24: Binary Classifier Accuracy for Inclusion and Min_Distance Combinations - Full, Continue

		90_2	**90_2.5**	**90_3**	**90_3.5**	**100_2**	**all**
Anger	LR	0.887437	0.896528	0.896508	0.889706	0.880173	0.851111
	LDA	0.903772	0.900586	0.89832	0.885605	0.887884	0.814334
	KNN	0.814782	0.806592	0.789817	0.803906	0.813418	0.795718
	CART	0.836592	0.827507	0.835212	0.836586	0.828422	0.817528
	NB	0.760333	0.761238	0.770321	0.776676	0.755784	0.756234
	SVM	0.779831	0.779831	0.779831	0.779831	0.779831	0.779831
	ABC	0.881061	0.882423	0.873355	0.87246	0.885605	0.875167
	Kmc	0.177818	0.105617	0.127236	0.075856	0.102939	0.109364
		90_2	**90_2.5**	**90_3**	**90_3.5**	**100_2**	**all**
Happiness	LR	0.851578	0.862937	0.86703	0.867026	0.845681	0.827962
	LDA	0.84115	0.857497	0.861573	0.869751	0.845245	0.733104
	KNN	0.847503	0.849313	0.841127	0.848398	0.846598	0.844784
	CART	0.787585	0.789404	0.776246	0.777164	0.77942	0.775302
	NB	0.586444	0.606882	0.601444	0.638182	0.597785	0.527425
	SVM	0.862028	0.862028	0.862028	0.862028	0.862028	0.862028
	ABC	0.844307	0.854307	0.859303	0.862941	0.85884	0.859755
	Kmc	0.084473	0.115197	0.097577	0.142077	0.106222	0.101183

		90_2	90_2.5	90_3	90_3.5	100_2	all
Neutral	LR	0.715856	0.719484	0.720852	0.706767	0.719044	0.705866
	LDA	0.694529	0.70042	0.714031	0.698143	0.691329	0.621446
	KNN	0.602812	0.584671	0.600078	0.661845	0.603723	0.581026
	CART	0.645041	0.64369	0.643659	0.654111	0.649998	0.666876
	NB	0.570171	0.571985	0.572896	0.564264	0.573789	0.585627
	SVM	0.614591	0.614591	0.614591	0.614595	0.614591	0.614591
	ABC	0.703566	0.718114	0.710854	0.714486	0.706314	0.705426
	Kmc	0.159823	0.133036	0.146606	0.117114	0.123383	0.10803
		90_2	**90_2.5**	**90_3**	**90_3.5**	**100_2**	**all**
Sadness	LR	0.825687	0.838863	0.835699	0.827982	0.8266	0.81298
	LDA	0.827057	0.83296	0.829342	0.81936	0.825695	0.727674
	KNN	0.698603	0.739914	0.72267	0.758523	0.698603	0.741715
	CART	0.781211	0.793965	0.764412	0.768515	0.773472	0.783052
	NB	0.747602	0.746242	0.738986	0.741248	0.75078	0.744412
	SVM	0.74355	0.74355	0.74355	0.74355	0.74355	0.74355
	ABC	0.827534	0.822522	0.812538	0.806652	0.826613	0.830243
	Kmc	0.13935	0.115358	0.105755	0.138034	0.092199	0.114329

Figure 7: Min_Distance vs. Inclusion of Features with Anger Characteristics

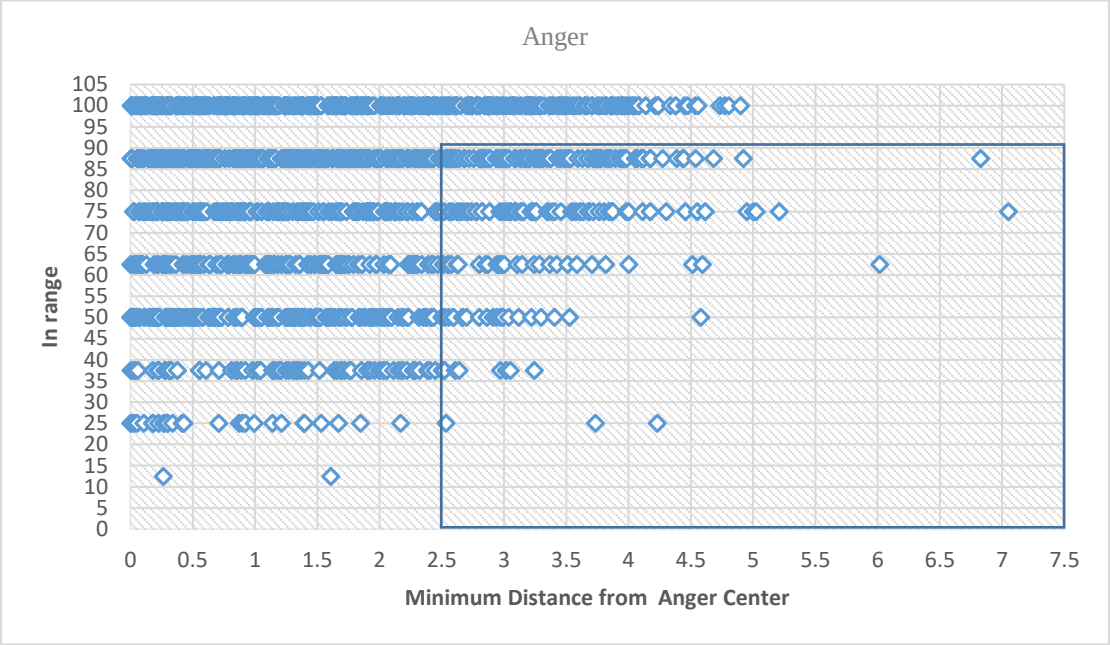


Figure 8: Min_Distance vs. Inclusion of Features with Neutral Characteristics

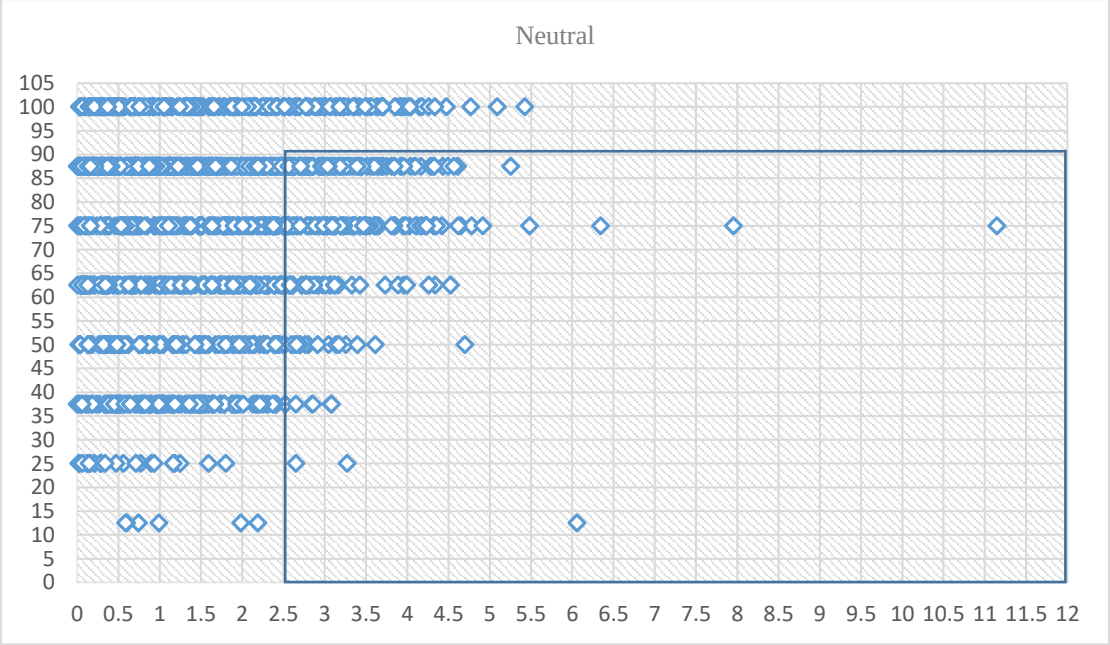


Figure 9: Min_Distance vs. Inclusion of Features with Sadness Characteristics

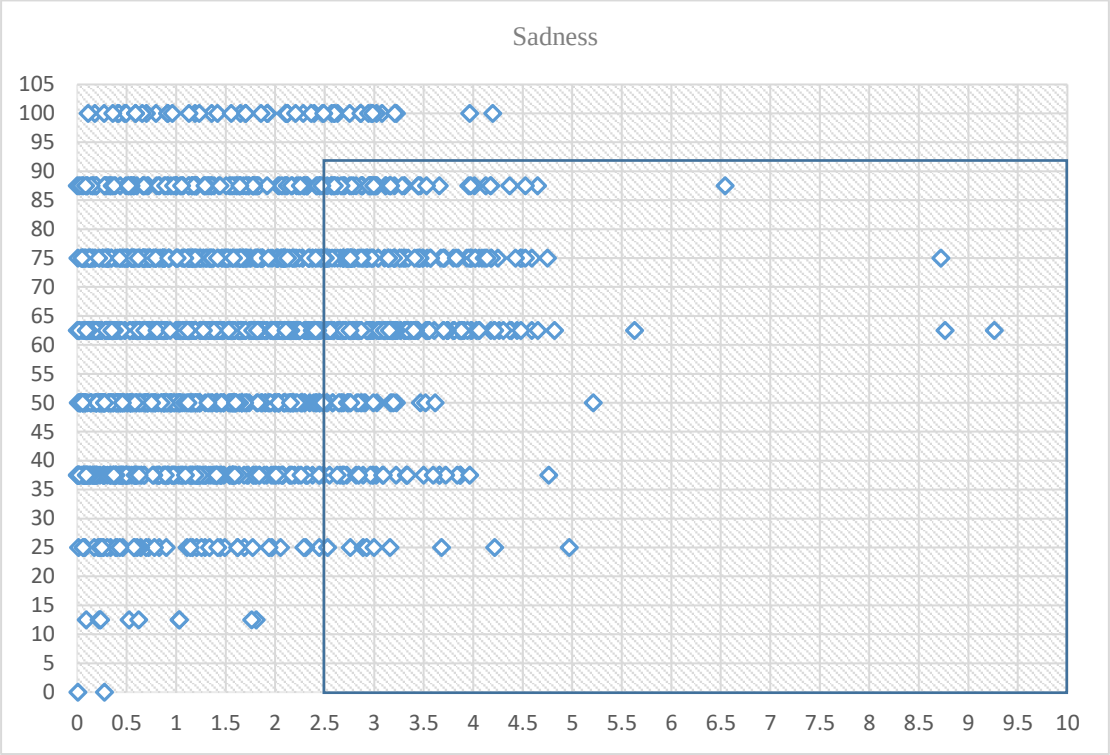


Table 25: Predicted Emotions in a Dyadic Conversation - Speaker 1

Utterance	Happiness	Sadness	Anger	Neutral
8	0	0	1	0
9	0	0	1	0
10	0	0	1	1
11	0	0	0	1
15	0	0	1	0
20	0	0	1	0
23	0	0	1	0
24	0	0	0	0
25	0	0	0	0
26	0	0	1	0
27	0	0	1	0
28	0	0	1	0
29	0	0	1	0
30	0	0	1	0
31	0	0	1	0
32	0	0	1	0
33	0	0	1	0
34	0	0	1	0
35	0	0	0	0
36	0	0	1	0
37	0	0	1	0
38	0	0	1	0
39	0	0	1	0
40	0	0	1	0
41	0	0	1	0
42	0	0	1	0
43	0	0	1	0
44	0	0	1	0
45	0	0	1	0

Table 26: Predicted Emotions in a Dyadic Conversation - Speaker 2

Utterance	Happiness	Sadness	Anger	Neutral
1	0	0	0	0
2	0	0	0	0
4	0	0	0	0
5	0	0	0	0
6	0	0	0	0
9	0	0	0	0
12	0	1	0	0
15	0	1	0	0
17	0	0	0	0
18	0	0	1	0
23	0	0	0	0
26	0	0	0	0
28	1	0	0	1
34	0	0	0	0
35	0	0	1	0
36	0	0	1	0
37	0	0	1	0
38	0	0	1	0
39	1	0	1	1
40	0	0	0	0
41	0	0	1	0
42	0	0	0	0
43	0	0	0	0
44	0	0	1	0