AJSE

American Journal of Science & Engineering

Volume 3 Issue 3

December 2022

American Journal of Science & Engineering (AJSE)
Society for Makers, Artists, Researchers and Technologists (SMART)
6408 Elizabeth Ave SE, Auburn 98092, Washington, USA
ISSN: 2687-9530 (Print) and 2687-9581 (Online)

Editor-in-Chief

Dr. Izzat Alsmadi

Texas A&M, San Antonio, USA

Research Interest: Cyber Intelligence, Cyber Security, Software Engineering, Social Networks

Bio: Izzat Alsmadi is an Assistant Professor in the department of computing and cyber security at the Texas A&M, San Antonio. He has his master and PhD in Software Engineering from North Dakota State University in 2006 and 2008. He has more than 100 conference and journal publications. His research interests include: Cyber intelligence, Cyber security, Software security, software engineering, software testing, social networks and software defined networking. He is lead author, editor in several books including: Springer The NICE Cyber Security Framework Cyber Security Intelligence and Analytics, 2019, Practical Information Security: A Competency-Based Education Course, 2018, Information Fusion for Cyber-Security Analytics (Studies in Computational Intelligence), 2016. The author is also a member of The National Initiative for Cybersecurity Education (NICE) group, which meets frequently to discuss enhancements on cyber security education at the national level.

Automatic Water Level Indicator in the Field Using Arduino

Mahmud Hasan
Department Of Computer Science Engineering
American Internation University Bangladesh
Email: mahmud.hasan9776@gmail.com

Sifat Rahman
Department Of Computer Science Engineering
American International University Bangladesh
Email: sifat491@gmail.com

Abstract— Water is a very important factor in agricultural production and is a key of our quality of life as well. Monitoring water level of a water source, such as dams, water tank or bore-well plays a key role in agricultural management and development. This work aimed at development of an automatic water level indicator. Developing a water level indicator can be helpful to reduce the water wastage. The major components used for the development of the water level indicator were Arduino board, sonar sensor, buzzer, LED lights. The automatic water level monitoring system was realized using a sonar sensor attached to an Arduino Uno to process the analog signal coming from the sensor into a useable digital value of distance also we added a buzzer in this system so that we can be aware about the increasing level of water.

Index Terms— automation, development, monitoring-system, water-level, sonar sensor.

I. INTRODUCTION

The project is Arduino based automatic water level controller and indicator. Here, we are going to measure the water level by using help of sonar sensors. We did this some researches to make this "Automatic water level indicator". By analyzing many resources, we find out what we need to use to build this. For developing the project, we got help through the internet. Our country is an agricultural country so this project will be helpful for our agricultural works to maintain the water level of our crop fields. There are many projects. Similar to this topic which motivated us to work on water level indicator. Also, we got the motivation by watching many real-life projects as well as we thought this project will be easy for us to make and also the water wastage problem in our country motivated us. Moreover, the changing weather or season which create a great impact in our crop field. It says that every crop requires different amount of water in different seasons and this can be done by using automatic water level controller which will also help in reducing wastage of water. This "Arduino based automatic water level indicator" will help us to protect our farmers' crop field from dryness and also from the excess water level which will aid to make our crops healthy as well. The system is basically designed to prevent the excessive amount of water in crop field. With the help of this technology water level can be controlled. A water pump will be there for the water source, a water exhaust motor and sensor to monitor the level of water in the crop field. Minimum level of water can be sensed by using a sensor and using a monitor water level can be observed. The monitor will notify whether the water level has increased or decreased. A buzzer will give alarm if the water level goes up. By using a generator extra amount of water can be reduced and can be provided if water level goes down. This project is structured as follows. In Section II we discussed the literature review to create a water level indicator. The methodology and modeling in Section III, results and discussion in Section IV and conclusion in Section V. References that we used to create this project and writing the report in Section VI.

II. LITERATURE REVIEW

In this paper, we use soil moisture sensor is used which is placed in the soil and water level sensor is used which is placed in the reservoir. The sensor sends the information to the microcontroller. An algorithm was designed which converts the analogue data of the sensor to percentage. This paper designed an automatic irrigation system controlled by a microcontroller ATMEGA328. The moisture sensor and the water level sensor send information to the microcontroller. When the moisture of the soil reduces below the set parameter the microcontroller automatically turns on the motor. The current soil condition is displayed in the LCD and any change in the state of the system is notified via the LED's, buzzer and the LCD display. Sandeep Kaur et al., [1] proposed an Automatic Irrigation System (AIS) for different crops with Wireless Sensor, Network (WSN) deploying sensor nodes in the agricultural field. Sensor nodes sense the soil temperature, sunlight, pH, relative humidity and groundwater parameters and different types of soil and crops at one time. Then sensor nodes send the sensed data to base station, where the data can be analyzed and meaningful data stored in the database and this data help the AIS in decision like: whether a crop requires water or not and the amount of water required by the plant. Although, the proposed AIS will reduce the wastage of water and save crop from unconditional seasons like rainfall condition and over irrigated and less irrigated conditions, but the drawbacks are numerous like if the base station is compromised, the entire system fails and sensor nodes are expensive. N. Siththtikumar et al., [2] prototyped a low-cost automated water irrigation system for home gardens using Arduino Uno, LCD, moisture sensors, solenoid valve, flow sensor and pipe lines. Moisture sensor array embedded in garden will sense the water level continuously, when water level goes low, the solenoid valve attached to the pipe line system will automatically open allowing water to flow to the garden via pipe line network. On the other hand, if the water level is sufficient enough the solenoid valve automatically closes restricting water flow to the garden. The LCD display will show the amount of water used in liters by sensing the water flow by the flow sensor also it shows the flow rate and temperature in the garden. The system is low cost and efficient for small garden but it needs improvement in other to act base on the soil type. Syed Musthak Ahmed et al., [3] proposes to make the farmers stay away from the field for the whole day, prototype that senses soil moisture and based on the data, the designed system instinctively turns ON the water pump to the field. As the soil reaches an optimum moisture level, then the water pump automatically gets turned OFF. It is shortly can termed as maintenance free agriculture where farmers can be prevented from breathing harmful chemicals by staying on the field for the whole day and also estimates the ambient temperature and humidity in the field and senses the daylight and rainfall intensity on the agricultural field. Divani et al., [4] elucidates that the major problem facing in modern society is the water resource shortage; agriculture is one of the demanding jobs which consumes large quantity of water. So, we must utilize maximum water in an efficient way. The components are moisture sensor; motor/pump and Arduino board are programmed using software. The predetermined range of soil moisture and temperature is set according to plant requirement. If soil moisture value is less than threshold system automatically triggers water pump on till sensor meets threshold and then sets off automatically. The value is passed on to the user network through an application. Kotni et al., [5] elucidates that the water content in the soil controls the action done by the Arduino. The soil moisture sensor will detect the water content in the soil and feed it to the data pin of the sensor and send the data to Arduino for further processing. The code used in this project focuses on the threshold moisture of the soil. If the data (moisture content) collected by sensor is greater than the threshold moisture required for the respective soil, then the Arduino gives blank feed to the motor enabling it to give blank output or none output. When the data (moisture content) is less than the threshold moisture required by the soil, the Arduino feeds the motor to pump the water from the sprinklers to the soil.

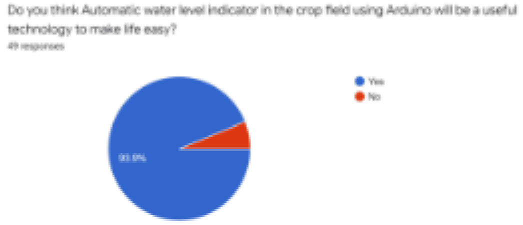
III. METHODOLOGY AND MODELING

3.1. Introduction

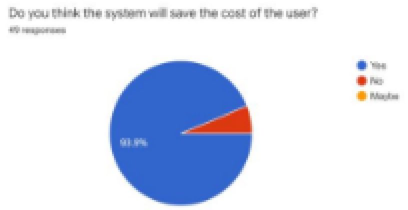
The Arduino Uno is utilized in this project. Many projects have been reported that employ the Arduino microcontroller, such as robots and mini-projects. The system is basically designed to prevent the excessive amount of water in crop field. With the help of this technology water level can be controlled. A water pump will be there for the water source, a water exhaust motor and sensor to monitor the level of water in the crop field.

3.2. Working principle of the proposed project

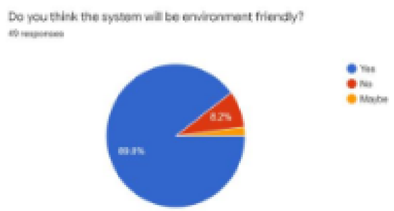
3.2.1. Process of Work



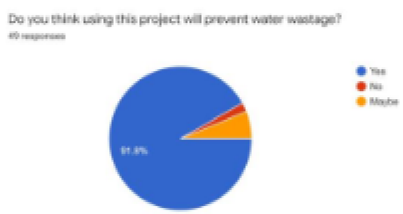
If we see the above pie chart it can be seen that 93.9% people agreed Automatic water level Indicator using Arduino will make life easier.



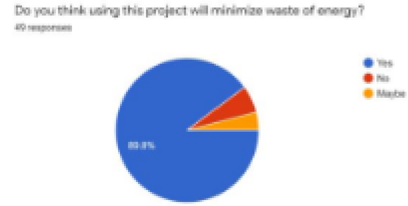
Cost friendly is really important and 93.9% of the people think this system will save the cost of the user.



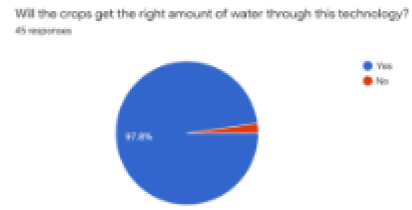
The system needs to be environment friendly and most of the people agreed to this statement as well which encouraged us to proceed to create a system like this using Arduino



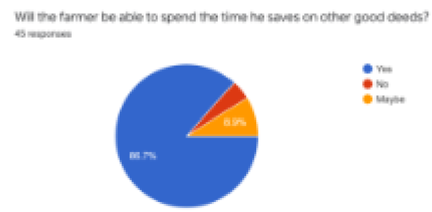
Saving water is very important for our country. 91.8% people agreed that this system might save the water and thus we moved further to do this project.



This project will definitely minimize the waste of energy and many people supported to this statement as well.



The main purpose of the project is to maintain the water level of crops so that they get the right amount of water and does not get damaged due to excess water.



Time saving is one of the main advantages of this system as the system is automated and here, we can see 86.7% of the people thought the same from this system.

3.3. Description of the important component

a) Arduino UNO: is an open source microcontroller board based on the Microchip ATmega328Pmicrocontroller and developed by Arduino.cc. The board is equipped with sets of digital and analog input/output (I/O) pins that may be interfaced with various expansion boards (shields) and other circuits. The board has 14 digital I/O pins (six capable of PWM output), 6 analog I/O pins, and is programmable with the Arduino IDE (Integrated Development Environment), via a type B USB cable. It can be powered by the USB cable or by an external 9-voltbattery, though it accepts voltages between 7 and 20 volts. It is similar to the Arduino Nano and Leonardo. The hardware reference design is distributed under a Creative Commons Attribution-Share-Alike 2.5 license and is available on the Arduino website. Layout and production files for some versions of the hardware are also available.

b) Grove Ultrasonic Ranger: is a non-contact distance measurement module which works at 40 KHz. When we provide a pulse trigger signal with more than 10uS through signal pin, the Grove Ultrasonic Ranger will issue 8 cycles of 40 kHz cycle level and detect the echo. The pulse width of the echo signal is proportional to the measured distance.

c) LED Light: A light-emitting diode (LED) is a semiconductor light source that emits light when current flows through it. Electrons in the semiconductor recombine with electron holes, releasing energy in the form of photons. The color of the light (corresponding to the energy of the photons) is determined by the energy required for electrons to cross the band-gap of the semiconductor. White light is obtained by using multiple semiconductors or a layer of light-emitting phosphor on the semiconductor device.

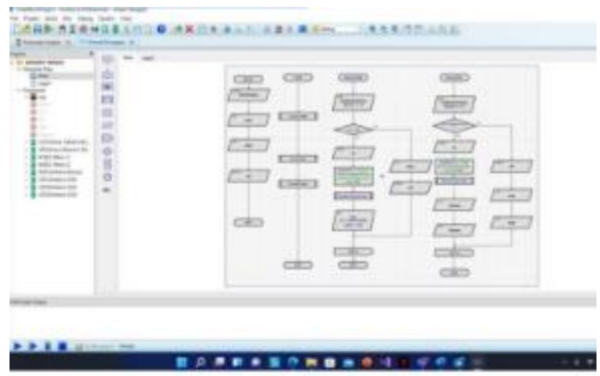
d) Buzzer: is used for given some indication and normally this indication is kind of a warning. Proteus has a built-in component for buzzer and it's an animated component means it gives a sound (beep) when it's turned ON.

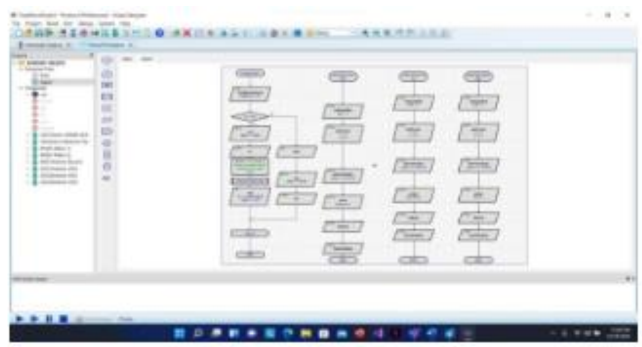
e) Motor Shield: is a driver module for motors that allows you to use Arduino to control the working speed and direction of the motor. Based on the Dual Full-Bridge Drive Chip L298, it is able to drive two DC motors or a step motor.

f) OLED Display: shows the messages which is written in the flowchart as per as conditions.

3.4. Implementation

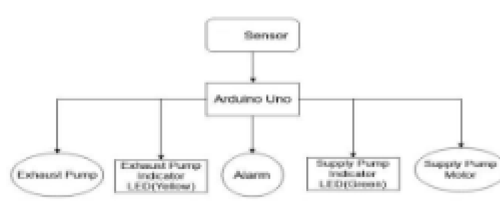
The system was created in order to maintain the water level of crops so that excess water does not harm the crops. The water level is ideal when the level is from 41cm to 50cm. The water level is considered as less when it is below 40cm and high when the level is above 50cm. When the ultrasonic ranger identifies the water, level is below 40cm, M1 starts at a forward direction at speed 255, LED1 turn on (green), OLED display shows "Supply pump is started and its supplying water to the Crop Field". When the ranger finds the water, range is between 40cm to 50cm the motors stay at ideal state and LED3 turns on (Yellow) and the display shows "Water level is at ideal situation for the Field" and both the motors are turned off automatically. When the ranger finds water, level is above 50cm, Buzzer starts, LED2(Red) turn on and M1v stops and M2 starts to rotate at a forward speed of 255 in order to pump out the excess water, display shows the message as "Exceeding pump is started and it's decreasing the water level."



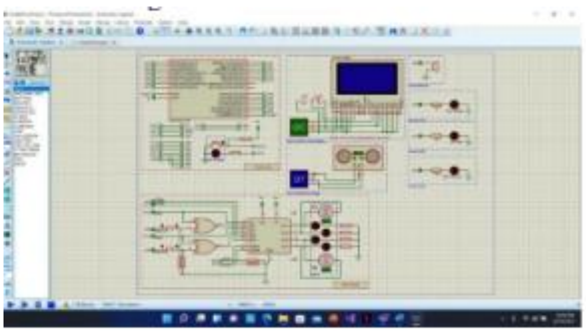


3.5. Test/Experimental setup

The primary design we used is from the block diagram. Initially we planned that Arduino Uno was programmed to control exhaust pump, alarm, supply pump and supply pump motor. The sensor was used to control Arduino Uno. Finally, we implemented the rough design in Proteus.



Finally, we implemented the rough design in Proteus. The challenge we faced was: at first, we thought to use Soil sensor but unfortunately, we did not get the required output so we used Grove Ultrasonic Ranger. The circuit shows how an automated water level indicator is designed and controlled using Arduino. The project is based on a microcontroller board design created by a number of suppliers utilizing a variety of microcontrollers. These systems come with a number of digital and analog input/output (I/O) pins that can be used to connect to expansion boards and other circuits. Arduino 328 was utilized in this project to program on what to do when a specified water level has reached and how to control the motors according to the water level. To begin, we create a schematic diagram



(As shown in Fig.) That will allow motors to run at various speeds. The motor's speed would be regulated by the Grove Ultrasonic Ranger, which will be achieved by adjusting the ranger's level, resulting all the motors to pump or exhaust water on at a specific speed. M1 motor is used to pump water into the fields and M2 is used to pump out the excess water. Arduino buzzer was used to alert if the water level rose higher than 50cm. In order to indicate different levels of water in the field 3 (Green, Yellow, Red) Arduino LED was used. LCD1 (Grove 128x64 OLED display) was used to show message when water is supplied, stopped & exceeded. The challenge we faced was: at first, we thought to use Soil sensor but unfortunately, we did not get the required output so we used Grove Ultrasonic Ranger.

3.6. Cost analysis

Each component has a different price. First and foremost, we must choose those that are significantly less expensive and hence more widely available. It must be conducted very successfully, despite the fact that it is less expensive; otherwise, it will fail. As a consequence, all of the chosen components fulfill our standards, and we can confidently state that it is both affordable and effective. The Arduino UNO R3 board cost 640 taka only. Grove 128x64 OLED display cost 550 taka. Dual channel motor shields cost 1,399 takas. Dc water pump motors cost 450*2 = 900 taka. Grove Ultrasonic Ranger cost 2,049 taka. Arduino buzzer cost 45 taka. We may simply implement this experiment within 6,583 BDT. We considered utilizing a Raspberry Pie Board for this project at first. However, it came at a cost of BDT 11,943 taka.

IV. RESULTS AND DISCUSSION

For this experiment we have used proteus software to build our project. In this software we have used Arduino Uno 328, Grove 128*64 OLED display, Grove Ultrasonic Ranger, Arduino Motor Shield(R3) with DC Motor, Arduino Buzzer, Arduino LED (green), Arduino LED (red) and Arduino LED (yellow).

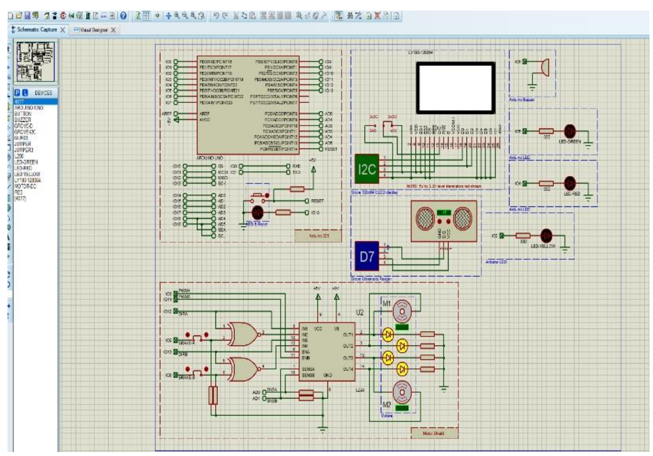


Figure 1 Schematic Capture of water level indicator in crop field

We have taken all these components from the peripheral which we can find in the flowchart Visual Designer section. From the Breakout Peripherals Category in Add peripheral section we have taken Arduino Buzzer, Arduino Led (Green), Arduino Led (Red), Arduino Led (Yellow). From the Grove Category we have taken Grove Ultrasonic Ranger

Module, Grove 128x64 OLED display Module. Also, from the Motor Control Category we have taken Arduino Motor Shield (R3) with DC Motors.

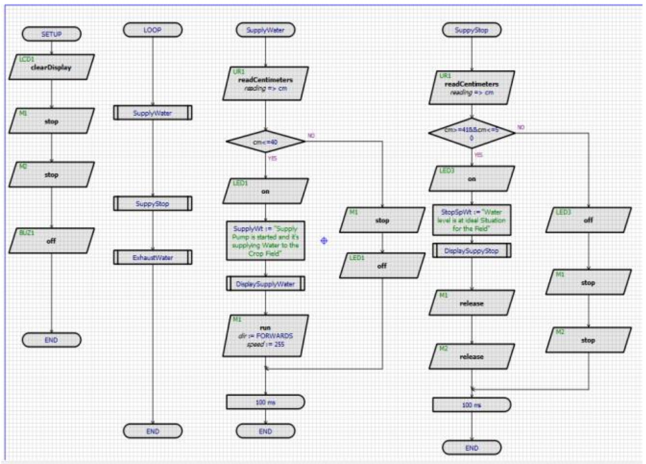


Figure 2 Flowchart of the Simulation

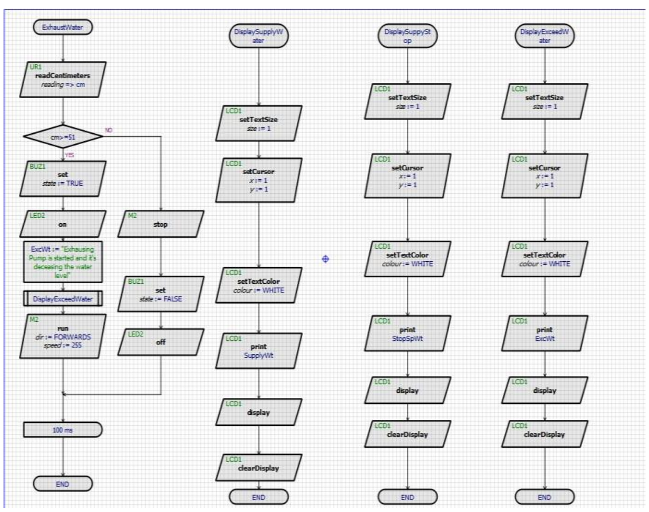


Figure 3 Flowchart of simulation page2

Here Figure 2 and 3 are the flowchart of the simulation.

There are 3 different modes in the water level indicator in the crop field. The simulation of some of the modes of this projects are given below.

Here in Figure 4, the M1 motor is running forward and M2 motor is stopped in that condition which defines that if the water level is less or equal 40 cm then the motor (Forward) will supply water and the Green Led light will be on.

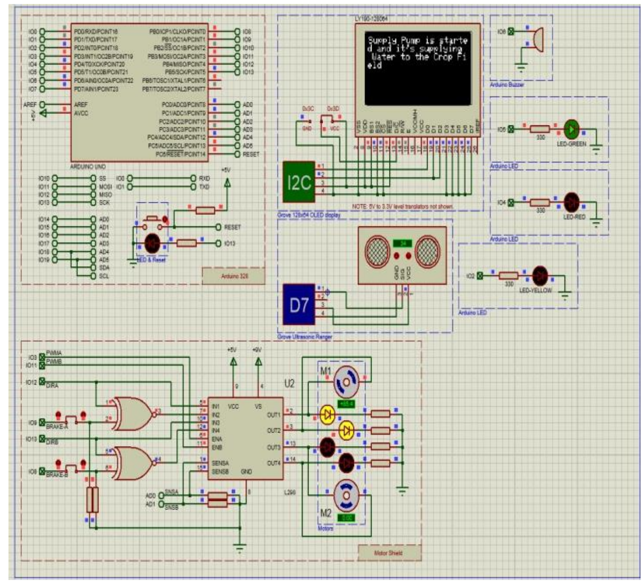


Figure 4 Water Supply Motor and Green Led is on.

Here in Figure 5, the water level is in ideal state and both of the motors are stopped. The water level of the ideal state is between 41 to 50 cm and led yellow light will be on.

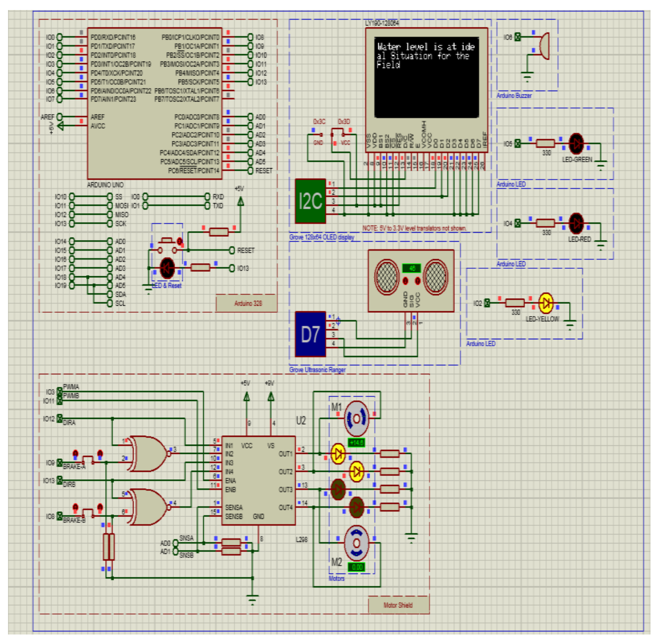


Figure 5 Both Motors are stopped and Yellow light is on.

Here in Figure 6, the water level is higher than ideal and M2 motor (Forward) will be on and it will exhaust the extra water. When the water level is higher or equal to 51 cm, it will be activated and the Buzzer will be on to alert about the situation and Red led light will be on.

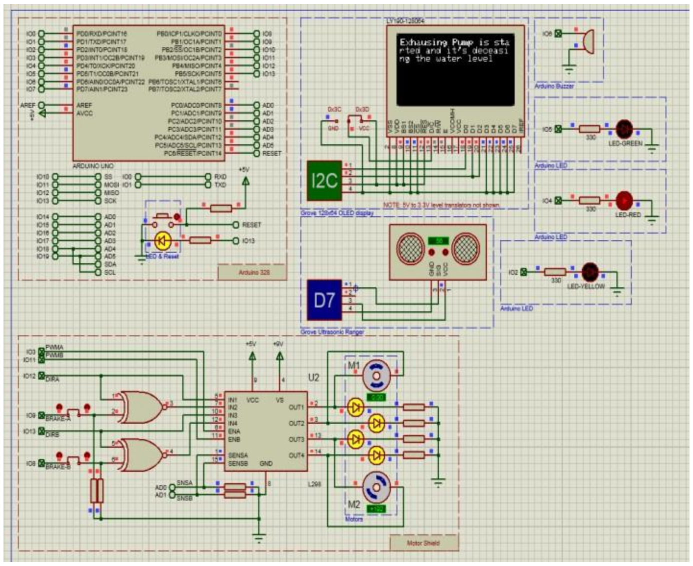


Figure 6 Exhaust motor, Arduino Buzzer and Red Led is on

V. CONCLUSION

In this experiment, the goal was to create a water level indicator for crop field. The speed of the motor was controlled by using a sonar sensor. The display was also used to monitor the different water level. We are using Arduino and the Proteus software in this project. After creating flowchart and build the project the simulation run successfully. With different percentage of dummy accelerometer value or potentiometer value, we had attained different results successfully. Therefore, the aim of this project was fulfilled. Due to time limitations, we couldn't able to add more features like buzzer which will give us alarm when the water level goes down and Xbee transmitter and receiver, which is a shield that allows Arduino to communicate wirelessly with the use of zigbee. This is basically, a prototype project that means when the industries will make it, they can make it at a minimal price. The main purpose of this project is to reduce the wastage of water and make a low-cost water level indicator system.

REFERENCES

1.Sandeep Kaur and Deepali, (2017) "An automatic irrigation system for different crops with WSN", 2017 6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Amity University Uttar Pradesh, Noida, India

2. N. Siththtikumar and M. W. P. Maduranga, "Designing and Implementing an Arduino Based Low-Cost Automated Water Irrigation System for Home Gardens", International Research Symposium on Engineering Advancements 2016 (IRSEA 2016)

3. Ahmed, S. M., Kovela, B., & Gunjan, V. K. (2020). IoT Based Automatic Plant Watering System Through Soil Moisture Sensing—A Technique to Support Farmers' Cultivation in Rural India. In Advances in Cybernetics, Cognition, and Machine Learning for Communication Technologies (pp. 259-268). Springer, Singapore.

4. Divani, D., Patil, P., & Punjabi, S. K. (2016, April). Automated plant Watering system. In 2016 International Conference on Computation of Power, Energy Information and Communication (ICCPEIC) (pp. 180-182). IEEE.

5. Siva, K. N., Bagubali, A., & Krishnan, K. V. (2019, March). Smart watering of plants. In 2019 International Conference on Vision Towards Emerging Trends in Communication andNetworking (ViTECoN) (pp. 1-4). IEEE.

Detecting Encryption Vulnerabilities with Lossless Compression

Richard Hansen
rhhansen@captechu.edu
Capitol Technology University
11301 Springfield Road, Laurel, MD 20708

ABSTRACT: Ciphertext entropy is a key property for detecting the use of insecure encryption modes such as Electronic Code Book (ECB) and for testing symmetric=key cryptographic algorithms. This paper discusses the use of lossless compression utilities as a proxy measurement for entropy and its use in detecting a limited set of encryption vulnerabilities. We also note the use of off-the-shelf algorithms yields a teaching tool for Information Technology and Cybersecurity students. Finally, small and consistent differences between encryption modes provide the potential to identify the encryption algorithm from compression factors.

Keywords: Cryptography, Cryptology, Encryption, Entropy, Compression, Cybersecurity, Vulnerability Assessment, Information Assurance, ECB

INTRODUCTION

In April, 2020 it was revealed that Zoom Meetings provided misleading information to clients regarding the encryption algorithm used for protection of confidentiality (Marczak and Railton, 2021). An encryption mode with known vulnerabilities, Electronic Code Book or ECB, was used without being disclosed to users and an unknown number of meetings were compromised/

This mode may be in use by other systems and applications. Developers, IT staff and the user community may not understand issues related to algorithms, modes and other encryption features (Bai et al., 2020). One potential method for detecting the use of ECB is measuring a property known as entropy for the output of the encryption process, the ciphertext. Entropy measurements can be useful for detecting a limited number of vulnerabilities including the use of ECB. Entropy measures the lack of order and predictability in a file; a low entropy measurement means there are patterns that can be detected or predicated and a high entropy measurement means there are few patterns that can be detected or predicted.

Compression is also used to identify patterns in data such as repeated characters or sequences of characters that have a mathematical relationship. The identified patterns are used to reduce the size of data so it will use less resources when stored or transmitted.

This paper details a qualitative study that provides the following contributions:

- Experimental results that validate the use of lossless data compression to detect the use of ECB mode in several algorithms, and detection of a limited number of issues related to poorly chosen encryption keys and initialization vectors.

- Experimental results that establish a relationship between measured entropy of ciphertext generated using symmetric key algorithms and the degree of lossless data compression possible with the commonly available GNU Zip utility.
- Techniques for use of lossless data compression as a teaching tool for the concepts of entropy as it relates to encryption/

The Problem – Detection of Vulnerable ECB Mode Encryption

Zoom Meetings claimed to use strong encryption to protect the confidentiality of meetings based on the Advanced Encryption Standard algorithm with a 256-bit key length (AES-256). In reality Zoom used a weaker AES variant, AES-128 in the Electronic Code Book (ECB) mode. ECB mode is vulnerable to exploitation when used for large amounts of data. Its use may be detected using specially selected inputs, known as plaintext, and entropy measurements of the resulting ciphertext output. Other problems may cause issues with encryption algorithms and their software implementations, such as values for keys and initialization vectors that are easily guessed or that produce undesired results.

Entropy refers to the lack of predictability in ciphertext. Ideally this means that knowledge of the previous bits and bytes in a file or stream of data does not allow prediction of the value of the next bits or bytes. Useful patterns should be obfuscated by a properly designed and implemented encryption algorithm which produces ciphertext with high degree of entropy (unpredictability).

Modern encryption algorithms such as the Advanced Encryption Standard (AES) use the principles of confusion and diffusion to obfuscate patterns in data and increase entropy. These terms originated with Claude Shannon’s then-classified paper “A Mathematical Theory of Cryptography” (Shannon, 1945). Confusion refers to representing a given pattern of bits in plaintext with a different pattern of bits in ciphertext. Diffusion refers to transposing bits, that is changing their position in a systematic manner. Applying both techniques increase the difficulty of detecting useful patterns in ciphertext.

Testing for the presence of low-entropy encrypted data may be difficult for IT and Cybersecurity staff. Applications are available to measure ciphertext entropy, however the knowledge and skills required for their use is not part the NIST NICE framework nor covered by the COMPTIA Security+ Exam (Newhouse, 2017). The next section of this paper details how the proposed method, data compression, may be used as a proxy measurement for entropy.

A literature search was performed to find related work. The paper Relationship Between Entropy and Test Data Compression (Balakrishnan & Touba, 2007), examines the performance of different compression techniques and for test data generated by system-on-a-chip designs. Entropy measurements are used to establish theoretical limits for the amount of compression. The authors’ detailed examination of compression algorithms is useful for those considering similar problems.

On Compression of Data Encrypted with Block Ciphers (Klinc et al., 2009) investigates the compression of high-entropy ciphertext. The authors discuss an approach may provide significant compression for certain sets of ciphertext inputs.

Distinguishing Compressed and Encrypted File Fragments (De Gaspari et al., 2020) examines the problem of using entropy to detect encrypted files when compressed files may have similar entropy values. Current approaches were not successful and the authors created a learning-based classifier, ExCoD that can differentiate between the two types of files.

COMPRESSION AS A PROXY FOR ENTROPY

Data compression is designed to decrease the amount of storage required for a given set of data by finding patterns that can be reduced to more compact representations. This led to the construction of a hypothesis for this paper and a supporting lemma:

- The author hypothesizes that it is possible to use compressibility as a proxy measurement of entropy for detecting the use of Electronic Code Book (ECB) mode encryption with a modern cryptographic algorithm and selected plaintext consisting of a single repeated character.
- The author proposes that there is mathematical relationship between the measured entropy of encrypted data and the measured amount of compression provided by readily-available applications.

An experiment was designed to generate data to test the hypothesis and the associated lemma. First, a series of plaintext files were specified and then generated. Then a process was designed to:

- Encrypt the plaintext files.
- Measure the entropy of the encrypted file.
- Compress the encrypted file.
- Calculate the change in size of the compressed file vs the uncompressed file.

A file size of 2,560 bytes was chosen. This will contain multiple blocks of data from the largest block size in use among modern cryptosystems. Plaintext files containing repeated single characters, nulls (0x00), were generated. Plaintext files containing a single set and multiple repeated sets of random binary data were also generated. The Linux "dd" utility was used to read random data from the Linux /dev/random device.

There are two types of file compression in general use, lossless and lossy. Lossless compression has the ability to exactly recreate the original data and is widely used for documents and other files that will suffer from changes to the data. Lossy compression provides a greater amount of compression at the cost of an inexact replication of the original data. Lossy compression is useful for video, images, audio, and other data where small differences are acceptable. Lossless compression was selected for use to allow for an exact recreation of the original files.

Two commonly available lossless algorithms are bzip2 and zip, both of which can be used from the command line for automating the encryption, compression and measurement process ("the process") on Windows and Linux. Experiments found that gzip encryption provides a greater amount of compression for ciphertext files and gzip was selected for use.

Data was encrypted using the "openssl" command-line encryption application. It supports many modern encryption algorithms and has command line options that assist with automating the process.

Entropy was measured using the Linux "ent" command-line application. The accuracy was measured against the Cryptool Window GUI application used for cryptography research and education. Results from the tools were compared and there was less than a 2% difference in measured entropy values. The "ent" application was selected for its ability to be used from the command line to automate the process.

The experiments and resulting data are described Experiments and Data below. Conclusions and Further Research describes the application of these techniques to Cybersecurity & Information Assurance education.

EXPERIMENTS AND DATA

Testing was performed using Kali Linux version 2021.2 in a VMWare virtual machine. Encryption was performed using the openssl utility version 1.1.1k. Entropy was measured using the ent utility, build date 11/22/20. Encryption, compression, and entropy measurements were automated with zsh shell scripts.

Input files were constructed to address the hypothesis and the associated lemma. A 2,560 byte file consisting of nulls (0x00, all bits set to "0") was used as plaintext. The use of a single 8-bit character negates concerns about byte-alignment within each block of data. Each algorithm and mode were tested with a key of all nulls, 3 randomly generated keys, and a key with all bits set to 0xFF repeated (all bits set to "1"). For those modes requiring Initialization Vectors (IVs), each algorithm and mode were tested with an IV containing all nulls, 3 random IVs, and an IV containing 0xFFs.

Three encryption algorithms and three modes of encryption were selected for testing. The first algorithm is the Data Encryption Standard (DES) which has a 56-bit key and a 64-bit block size. DES was originally developed by IBM and is the oldest of the three algorithms. The second algorithm is the SEED algorithm which has a 128 bit-key and a 128-bit block size. SEED was developed by the South Korean government for use by South Korean government, defense, and commercial organizations. The third algorithm, the Advanced Encryption Standard (AES), is a modern symmetric cipher endorsed by the United States' National Institute of Standards and Technology (NIST). AES can use a 128-bit, 192-bit, or 256-bit key and has a 128-bit block size. The 256-bit key size was used for these tests.

Each algorithm was tested in its Electronic Code Book (ECB), Cipher Block Chaining (CBC), and Output Feed Back (OFB) modes. For any given algorithm and key, ECB provides the same result each time a given piece of data is encrypted using that key. CBC and OFB are more secure modes that use an initialization vector (IV) in addition to the key.

Plaintext files containing sequences of random characters were included to measure their effect on entropy and compression. The first file contained 2,560 bytes of random characters (0x00-0xFF), the second contained 5 repeated sections of 512 random bytes, the third contained 10 repeated sections of 256 random bytes, and the fourth contained 20 repeated sections of 128 random bytes.

During testing it was noted that the size of the ciphertext filename had a small effect on the resulting file size due to large filenames requiring more storage space than small filenames. The calculations for compression percentages made accommodations for this issue.

The input files, shell scripts used to automate the process, and files containing experimental results are have been uploaded to Github and are available to the public.

Proving the Hypothesis

Figure 1 below shows the compressed size of the ciphertext output for the ECB algorithm as a percentage of its original size, the type of plaintext, and the encryption key. In all cases encryption of plaintext consisting of nulls resulted in ciphertext that was compressed to a small fraction (2%-3%) of its original file size.

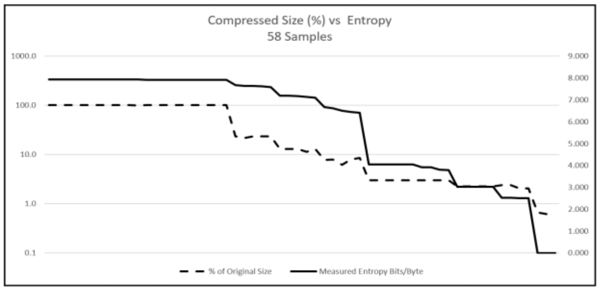


Figure 1 Compressed Size of Ciphertext vs

Measured Entropy

Note - (R) =Random				Compressed
Algorithm	Plaintext	Key	Entropy	Size
DESECB	Nulls	Key 0 - Nulls	3.030	2.26%
DESECB	Nulls	Key 1 - (R)	3.027	2.26%
DESECB	Nulls	Key 2 - (R)	3.030	2.26%
DESECB	Nulls	Key 3 - (R)	3.030	2.26%
DESECB	Nulls	Key 4 - 0xff's	3.030	2.26%
DESECB	Random (R)	Key 0 - Nulls	7.942	101.25%
DESECB	5x512 (R)	Key 0 - Nulls	7.637	23.44%
DESECB	10x256 (R)	Key 0 - Nulls	7.187	12.93%
DESECB	20x128 (R)	Key 0 - Nulls	6.678	7.75%
DESECB	Random (R)	Key 1 - (R)	7.933	101.25%
DESECB	5x512 (R)	Key 1 - (R)	7.577	23.40%
DESECB	10x256 (R)	Key 1 - (R)	7.106	12.97%
DESECB	20x128 (R)	Key 1 - (R)	6.406	8.57%
SEED128ECB	Nulls	Key 0 - Nulls	4.054	3.03%
SEED128ECB	Nulls	Key 1 - (R)	4.050	3.03%
SEED128ECB	Nulls	Key 2 - (R)	3.930	3.03%
SEED128ECB	Nulls	Key 3 - (R)	3.800	3.03%
SEED128ECB	Nulls	Key 4 - 0xff's	4.050	3.03%
SEED128ECB	Random (R)	Key 0 - Nulls	7.922	101.32%
SEED128ECB	5x512 (R)	Key 0 - Nulls	7.670	23.56%
SEED128ECB	10x256 (R)	Key 0 - Nulls	7.188	13.12%
SEED128ECB	20x128 (R)	Key 0 - Nulls	6.613	7.92%
AES256ECB	Nulls	Key 0 - Nulls	3.803	2.99%
AES256ECB	Nulls	Key 1 - (R)	4.047	2.99%
AES256ECB	Nulls	Key 2 - (R)	4.048	2.99%
AES256ECB	Nulls	Key 3 - (R)	4.054	2.99%
AES256ECB	Nulls	Key 4 - Nulls	3.923	2.99%
AES256ECB	Random (R)	Key 0 - Nulls	7.926	101.28%
AES256ECB	5x512 (R)	Key 0 - Nulls	7.623	23.45%
AES256ECB	10x256 (R)	Key 0 - Nulls	7.168	13.04%
AES256ECB	20x128 (R)	Key 0 - Nulls	6.453	7.88%

Figure 2 – ECB Compression Results

Encrypting a totally random plaintext file results in a file that is a approximately 1% larger than the original file for all algorithms. The larger size is due to gzip compression's internal representation of data that has few useful patterns in the ciphertext; the original file is replicated with additional overhead needed by the gzip algorithm. The table also shows that compressing files with repeated random sequences resulted in significant decreases in size and lower values for entropy.

The results shown in Figure 1 and 2 above prove the Hypothesis. Lossless compression can be used as a proxy for entropy to detect the use of ECB using chosen plaintext for the three algorithms tested.

Proving the Lemma

As shown in figures 1 and 2 above, compression and entropy were measured for 58 ciphertext and plaintext files. Entropy measurements ranged from 0.00 up to 7.947 where 0.00 is the minimum possible and 8.00 is the maximum possible. The compressed files ranged from 0.6% to 101.2% of the uncompressed file size.

The graph in Figure 1 shows that a logarithmic relationship exists between measured entropy (0-8) and the compressed file size expressed as a percentage of its original size, proving the lemma.

Detecting Other Vulnerabilities, DES

"Weak Keys"

Some implementations of DES are known to have issues with certain keys, such as repeated 0x00 (nulls, no bits set) and repeated 0xff (all bits set). These resulted in highly compressed files (2% of original size) and they had lower entropy than files created with random keys (2.5 vs 7.9) as shown in Figure 3 below.

Note - (R) =Random					Compressed
Algorithm	Plaintext	Key	IV	Entropy	Size
DES CBC	Nulls	Key 0 - Nulls	IV 0 - Nulls	2.532	2.38%
DES CBC	Nulls	Key 1 - (R)	IV 1 - (R)	7.923	100.97%
DES CBC	Nulls	Key 2 - (R)	IV 2 - (R)	7.915	100.97%
DES CBC	Nulls	Key 3 - (R)	IV 3 - (R)	7.936	100.97%
DES CBC	Nulls	Key 4 - 0xff's	IV 4 - 0xff's	2.532	2.41%
DES OFB	Nulls	Key 0 - Nulls	IV 0 - Nulls	2.500	2.07%
DES OFB	Nulls	Key 1 - (R)	IV 1 - (R)	7.923	100.98%
DES OFB	Nulls	Key 2 - (R)	IV 2 - (R)	7.915	100.98%
DES OFB	Nulls	Key 3 - (R)	IV 3 - (R)	7.937	100.98%
DES OFB	Nulls	Key 4 - 0xff's	IV 4 - 0xff's	2.500	2.07%

Figure 3 - Key and IV Issues with

OPENSSL DES Encryption

These results were unexpected by the researcher. Further investigation showed that this vulnerability and the underlying implementation issues are well known.

SEED and AES – Non-ECB Modes

The SEED and AES encryption algorithms had a compressed file size slightly larger than the ciphertext files themselves, indicating little or no compression was possible when used the CBC and OFB modes. The output files had a consistently high entropy in excess of 7.92 as shown in Figure 4 below.

Note - (R) =Random					Compressed
Algorithm	Plaintext	Key	IV	Entropy	Size
SEED128CBC	Nulls	Key 0 - Nulls	IV 0 - Nulls	7.934	101.05%
SEED128CBC	Nulls	Key 1 - (R)	IV 1 - (R)	7.926	101.05%
SEED128CBC	Nulls	Key 2 - (R)	IV 2 - (R)	7.936	101.05%
SEED128CBC	Nulls	Key 3 - (R)	IV 3 - (R)	7.947	101.05%
SEED128CBC	Nulls	Key 4 - 0xff's	IV 4 - 0xff's	7.928	101.05%
SEED128OFB	Nulls	Key 0 - Nulls	IV 0 - Nulls	7.934	101.05%
SEED128OFB	Nulls	Key 1 - (R)	IV 1 - (R)	7.925	101.05%
SEED128OFB	Nulls	Key 2 - (R)	IV 2 - (R)	7.935	101.05%
SEED128OFB	Nulls	Key 3 - (R)	IV 3 - (R)	7.946	101.05%
SEED128OFB	Nulls	Key 4 - 0xff's	IV 4 - 0xff's	7.927	101.05%
AES256CBC	Nulls	Key 0 - Nulls	IV 0 - Nulls	7.930	101.01%
AES256CBC	Nulls	Key 1 - (R)	IV 1 - (R)	7.932	101.01%
AES256CBC	Nulls	Key 2 - (R)	IV 2 - (R)	7.932	101.01%
AES256CBC	Nulls	Key 3 - (R)	IV 3 - (R)	7.931	101.01%
AES256CBC	Nulls	Key 4 - 0xff's	IV 4 - 0xff's	7.931	101.01%
AES256OFB	Nulls	Key 0 - Nulls	IV 0 - Nulls	7.930	101.02%
AES256OFB	Nulls	Key 1 - (R)	IV 1 - (R)	7.931	101.02%
AES256OFB	Nulls	Key 2 - (R)	IV 2 - (R)	7.933	101.02%
AES256OFB	Nulls	Key 3 - (R)	IV 3 - (R)	7.930	101.02%
AES256OFB	Nulls	Key 4 - 0xff's	IV 4 - 0xff's	7.930	101.02%

Figure 4 High-Entropy – SEED and AES in

CBC and OFB Modes

Entropy measurements and the size of compressed files indicate that these algorithms successfully transformed plaintext with very low entropy, 0.0, into very high entropy ciphertext. As shown in Figure 4, a possible indicator of high-entropy ciphertext is an expansion in size when compressed with GNU Zip.

OTHER APPLICATIONS

The use of compression as a proxy measurement for entropy would be useful for educating professionals in the workplace and students through hands-on methodologies such as Discovery Learning and Experiential Learning. Exercises based on compressed file size may be used to provide competency at Levels 1, 2 and 3 of Bloom's Revised Taxonomy (Remembering, Understanding, Applying) (Krathwohl, 2002, p. 215). Comparisons can be made using easily understood metrics such as file sizes pre- and post-compression. Identification of repeated sequences can be related to compression, where repeated sequences are stored once and compact "pointers" act as placeholders for the sequences in the compressed file.

Jerome Bruner's Theory of Discovery proposes that learners use past experiences and knowledge to discover new facts and relationships through hands-on interaction with their environment (Mcleod, 1970). Experiential Learning also emphasizes hands-on experimentation. Lab experiments that provide for rapid feedback from experimentation will help students quickly build an internal representation for the relation between compressibility and entropy. Per Kolb, "the methods of grasping experience are abstract conceptualization and concrete experience "(Kolb, 1984)(Cherry, 2020). Conventional curriculums based on mathematics is focused on abstract representations. A complementary or alternate approach can use hands-on experience with the learner performing tasks and observing results in real-time to build their own base of knowledge and experience. Students will then be better prepared to understand abstract concepts such as entropy and Shannon's work on Information Theory.

CONCLUSIONS AND FURTHER RESEARCH

Conclusions

The experimental results support the conclusion that ECB-mode encryption for symmetric algorithms such as DES, SEED, and AES can be detected via use of chosen plaintext input and then compressing the ciphertext output with the GNU Zip lossless compression application. This conclusion may be generalized due to ECB-mode's inherently vulnerability to this type of analysis, and because a number of file compression algorithms and applications would be able to efficiently compress this type of data.

The experimental results indicate there is a logarithmic relationship between the compressed file size and the measured entropy of ciphertext files encrypted using the selected modern symmetric-key algorithms, proving the lemma. This relationship was also noted in the plaintext files used for encryption.

An unexpected finding is the potential to identify the encryption algorithm (DES, SEED, AES, etc.) using chosen ciphertext and the amount of compression. Small and consistent differences are noted between different algorithms and encryption modes. Non-ECB modes for s returned a consistent size of 101.98% compressed, SEED 101.05% compressed, and AES 101.01% compressed. An opportunity for further research would test other algorithms and key sizes (Triple-DES, AES-128, Blowfish, etc.).

Further Research

Additional research opportunities include determining if transmitted data has byte-alignment or encoding issues that make detection of ECB more difficult when data is distributed across packets for transmission. It is also possible that lossy compression may be useful for finding patterns that are not detected with lossless compression. A comparison between the two types of compression would be useful.

ACKNOWLEDGEMENTS

Capitol Technology University was kind enough to allow the use of its campus and facilities for this research. The late Dr. Win Wenger provided valuable guidance on the works of Drs. Piaget, Bruner & Kolb. Dr. William Butler, Dr. Sandy Antunes and Professor Frank Davis provided much helpful input and encouragement.

REFERENCES

Bai, W., Pearson, M., Kelley, P. G., & Mazurek, M. L. M. L. (2020, October 22). *Improving non-experts' understanding of end-to-end encryption*. EUSEC20. Retrieved November 20, 2021, from https://eusec20.cs.uchicago.edu/eusec20-Bai.pdf

Balakrishnan, K. J., & Touba, N. A. (2007). Relationship between entropy and Test Data Compression. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 26*(2), 386–395.

https://doi.org/10.1109/tcad.2006.882600

Cherry, K. (2020, May 15). *The David Kolb theory of how experience influences learning*. Verywell Mind. Retrieved March 22, 2022, from https://www.verywellmind.com/experiential-learning-2795154

De Gaspari, F., Pagnotta, D. H. G., De Carli, L., & Mancini, L. V. (2020, October 15). *Encod: Distinguishing compressed and encrypted File Fragments*. arxiv.org. Retrieved October 13, 2021, from https://arxiv.org/pdf/2010.07754.pdf

Klinc, D., Hazay, C., Jagmohan, A., Krawczyk, H., & Rabin, T. (2009). On compression of data encrypted with block ciphers. *2009 Data Compression Conference*. https://doi.org/10.1109/dcc.2009.71

Kolb, D. A., & Kolb. (1984). *Experiential learning: Experience as the source of learning and development, 2nd edition*. Pearson. Retrieved March 20, 2022, from https://www.pearson.com/us/higher-education/program/Kolb-Experiential-Learning-Experience-as-the-Source-of-Learning-and-Development-2nd-Edition/PGM183903.html

Krathwohl, D. R. (2002). A revision of Bloom's Taxonomy: An Overview. *Theory Into Practice*, *41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Marczak, B., & Scott-Railton, J. (2021, June 29). *Move fast and roll your own crypto: A quick look at the confidentiality of zoom meetings*. The Citizen Lab. Retrieved December 8, 2021, from https://citizenlab.ca/2020/04/move-fast-roll-your-own-crypto-a-quick-look-at-the-confidentiality-of-zoom-meetings/

Mcleod, S. (1970, January 1). *[bruner - learning theory in education]*. Simply Psychology. Retrieved June 9, 2021, from https://www.simplypsychology.org/bruner.html

Newhouse, W., Keith, S., & Scribner, B. (2017, August). National Initiative for Cybersecurity Edcation NIST 800-181, KSA K0274

Shannon, C. (n.d.). *A Mathematical Theory of Cryptography*. International Association for Cryptologic Research. Retrieved October 10, 2021, from https://www.iacr.org/museum/shannon/shannon45.pdf

Sentiment Analysis on COVID-19 Twitter Data

Srestha Sadhu[1]
Department of Computer Scienceand Engineering
University of Engineering andManagement, Kolkata.
srestha.sadhu@gmail.com

Varsha Poddar[2]
Department of Computer Scienceand Engineering
University of Engineering andManagement, Kolkata.
varsha.poddar@gmail.com

Puja Paul[3]
Department of Computer Scienceand Technology
University of Engineering andManagement, Kolkata.
puja.paul.uemk.cst.2023@gmail.com

Sheraly Hansda[4]
Department of Computer Scienceand Technology
University of Engineering andManagement, Kolkata.
sheraly.hansda.uemk.cs.2023@gmail.com

Rimpa Saha[5]
Department of ComputerScience and Technology
University of Engineering and Management, Kolkata.
rimpa.saha.uemk.cs2023@gmail.com

Angira Chakraborty[6]
Department of Computer Science andEngineering
University of Engineering andManagement, Kolkata.
angira.chakraborty.uemk.cse.2023@gmail.com

Titiksha Paul[7]
Department of Computer Science and EngineeringUniversity of Engineering and Management, Kolkata.
titiksha.paul.uemk.cs.2023@gmail.com

Anushree Mondal[8]
Department of Computer Science and InformationTechnology
University of Engineering and Management, Kolkata.
anushree.mondal.uemk.csit.2023@gmail.com

***Abstract*— Coronavirus first appeared in December 2019 in Wuhan, China which eventually lead to a catastrophic impact allover the world. The entire world had been fighting this pandemic and expressing their feelings, sharing their opinions on various social media platforms. Substantially Twitter had been the better medium to express their opinion and share updates about the situation. This paper analyzes the sentiments of the public based on positive,negative, and neutral tweets. This analysis eventually helps in the prediction of the covid-19 situation in the world. The dataset was collected from Kaggle which was uploaded by Gabriel Prada containing more than 1,70,000 tweets. Based on these tweets posted on Twitter a sentiment analysis was performed. Data Collecting, Data cleaning (Removing URL, #, @ and various types of punctuation), Tokenization, Stemming, removing stop- words were performed, and to find the polarity, two types of analyzers were used that is TextBlob and Afinn. 1,79,108 tweets were manually analyzed and comparing it with both the analyzers shows Afinn is more accurate.To evaluate the accuracy a few Machine learning Algorithms had been applied (Logistic Regression, Naive Bayes, Decision Tree, and Linear Regression) for predicting the sentiment of the tweet.**

***Keywords*— *COVID-19, Sentiment analyzers, Logistic Regression, Decision Tree, NaïveBayes, Linear Regression*.**

I. INTRODUCTION

Social media platforms such as Facebook, Twitter, and YouTube, provide us with information known as social data. This data is used for analyzing and predicting the future of various fields.Twitter had been chosen for this research as the platform to convey the thoughts and emotions of people worldwide. Twitter, a social networking and an online news site is used to communicate in short messages called tweets. Coronavirus known as COVID-19 is one of themost recently discussed topics in the world which has impacted negatively the day-to-day life of people. Many have lost their dear ones, lost their jobs, traveling is still restricted at some parts of the world and businesses are running at losses. Hence people chose to share their feelings in the form of help, gratitude, positively and expressed their pain with the world on social media platforms since the beginning of the pandemic, March 2020. Followed by the current situation this work focuses on the sentiments of people by collecting tweets from an already available dataset taken from Kaggle uploaded by Gabriel Preda. This dataset consists of data, dating from 25.7.2020 to 29.8.2020. Based on the positive, negative, and neutral tweets the impact of the coronavirus on the minds of people had been analyzed.

Sentiment Analysis on "COVID-19" Twitter data had been performed by using variousAnalyzers and Machine Learning Algorithms to find the polarity as well as the accuracy of the following Twitter dataset. Python had been chosen since it provides numerous libraries to access social media platforms like Twitter.

The dataset has been collected and cleaned through data cleaning, tokenization, stemming, and removing stop-words. For performing sentiment analysis both TextBlob and Afinn analyzer had been used to evaluate the polarity and then compared both the results. It has been found that using TextBlob the accuracy is more compared to Afinn but after evaluating the sentimentsof each tweet manually it was observed that Afinn is more precise. To find the accuracy, various machine learning algorithms such as Logistic Regression, Naïve Bayes, Decision Tree, and Linear Regression were used.

The results showed more neutral and positive tweets than negative tweets. Based on this, the maximum accuracy of 96.65% using the Logistic Regression algorithm was obtained whereas the other algorithms gave 82.09%, 95.51%, and 70.99% respectively.

The main objective is to analyze the sentiment of people due to covid-19 between July, 2020- August, 2020. During this period, coronavirus was in its initial phase and the Centers for Disease Control and Prevention (CDC) were ongoing research for the vaccine to develop and manufacture them. With rapid research development, people were expecting that the world would come up with a solution for the situations to get better globally. Hence, people were spreading awareness worldwide and holding campaigns for free RT-PCR tests. Therefore, it is expected to get more neutral and positive tweets than negative tweets.

II. RELATED WORKS

Many researchers have performed elaborate studies to show the effectiveness of Machine learning in sentiment analysis of people throughout the world fighting with the Virus. COVID-19 has caused a global crisis which has changed the perception of the world and compelled people to deal with the large scale disaster caused by it which has also impacted people, psychologically. Data from Twitter contributes and helps to discover sentiments of the people in various cases when the world is facing a pandemic situation .

This section of the paper covers several important papers related to COVID-19 sentiment analysis dated from 2020 to 2021 which were used as references. Researchers of [1] [4] [5] [6][7] [8] [11] have performed Public Sentiment Analysis on Twitter Data since the outbreak of COVID-19 in 2020 where preprocessing is done using NLTK library and TextBlob analyzer is used to analyze the polarity and subjectivity of tweets.

A survey conducted by Abdul Aziz et al.[2] in 2021, indicated the use of Naïve Bayes Classifier (NBC) and Logistic Regression (LR) classification methods of machine learning. In 2020, Stanislaw Wrycza [3] also published a research work using Naïve Bayes. In 2020, Drias et al.[9] performed a study using the lexicon-based approach. In 2021, Gupta et al.[10] published research work to classify the data accurately. Eight different classifiers are used (Multinomial Naive Bayes, Bernoulli Naive Bayes, Logistic Regression, LinearSVC, AdaBoost Classifier, Ridge Classifier, Passive Aggressive Classifier, and Perceptron).

In 2021, Supriya Raheja[12] published a research work that visualizes textual data through WordCloud and categorizes Tweets into notions like positive, negative, and neutral. In 2021, Piyush Ghasiya [13] published a study by creating a labeled dataset using unsupervised machine learning methods. In 2020, Machuca et al. [14] performed research work aiming to deduce whether the sentiment is positive or negative by applying machine learning algorithmsand NLP techniques. In 2021, Amir Hussain [15] worked on a research paper where social media data is largely unstructured, but natural language processing (NLP) and machinelearning (ML) is used. In 2021, Carol Shofiya [16] used the Hybrid Approach by employing the sentiment polarity and then applying the SVM algorithm, for classification and analysis. In 2021, Amir Rehman [17] disseminated a research work using CNN and D.N.N. which were the most suitable classification techniques to detect sentiment, followed by SVM, Random Forest, K-NN, and L.S.T.M.

By going through all the papers it is found that TextBlob analyzer and Naïve Bayes algorithmare used the most. By gathering information from all these papers and keeping them as references, we are inspired to proceed further with our study. The organization of the paper is as follows: Section 3 describes the subject dataset while Section 4 describes elaborately the implementation mechanism under which the data preprocessing and application model is explained. In Section 5, the Results and Findings of the research are given elaborately. Section6 and Section 7 give the conclusion and future scope of the paper, respectively.

III. SUBJECT DATASET

In this research, one subject dataset with 13 attributes has been chosen upon which all the necessary operations have been implemented. This dataset has data records from 25/7/2020 to 29/8/2020. In this period, 179108 tweets have been tweeted by the people regarding the COVID-19 pandemic. According to the graph of that time, corona cases were less than other times. Our work is to find out the sentiment of people over this period.

IV. IMPLEMENTATION MECHANISM

A. Data preprocessing:

The prime focus of pre-processing is to clean the data and prepare for further implementations.Firstly all the tweets have been cleaned by removing contains like hashtags, user handles, whitespaces, URLs, punctuations, and stop-words. Then tokenization and stemming have been done on the cleaned tweet.

Now two analyzers named as TextBlob and Afinn were applied to estimate the polarity scoresas well as the sentiment of each tweet in the dataset. The cleaned tweets from the previous stepwere subjected to multiple evaluation models using TextBlob and AFINN.

TextBlob library supports complex analysis and operations on textual data. Here it is used to get polarity and subjectivity scores associated with each tweet.

AFINN is one of the most popular lexicon-based approaches used for sentiment analysis whichis suitable for many languages. It contains a method called score() which takes a sentence as input and returns polarity score as output. It also contains more than 3300 words with a polarityscore associated with each word. A generalized score of polarity was found for each tweet. Both analyzers return a value in the range [-1to+1] where +1 implies Extreme Positive Polarity,-1 implies Extreme Negative Polarity and 0 implies neutral.

Finally, the dataset was ready to split into the form of training and testing data. Here 80% of the record was selected as train data while the rest 20% was considered as test data.

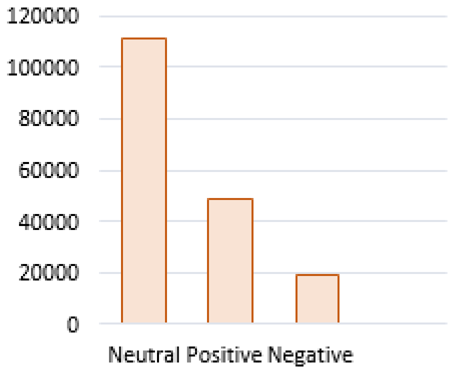


Fig 1. Using TextBlob Analyzer

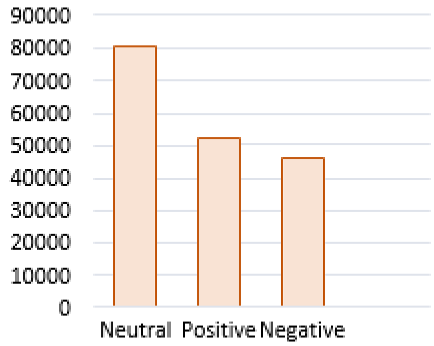


Fig 2. Using Afinn Analyzer

Fig 1 shows the diagrammatical representation of the number of neutral tweets that is 111110,positive tweets are 48904 and negative tweets are 19094.

Fig 2 shows the diagrammatical representation of the number of neutral tweets that is 80798, positive tweets are 52433, and negative tweets are 45877.

It has been found that using TextBlob analyzer the number of neutral tweets is more comparedto Afinn but after evaluating the sentiments of each tweet manually it was observed that Afinn is more precise and accurate therefore TextBlob analyzer has been discarded for further application.

B. Application of Models:

In this research, for predicting the sentiment of the tweets various supervised machine learning algorithms were implemented:

Logistic Regression: Logistic Regression is a binary classification algorithm, used to predict a dependent variable based on a set of independent variables such that the dependent variable is categorical. The name is "regression" because its working technique is quite similar to linear regression.

Decision Tree: It is a supervised machine learning algorithm used to solve both classification and regression problems. It is a tree that helps us in decision-making purposes.

Naïve Bayes: It is a supervised machine learning algorithm based on the Bayes theorem used for solving classifications problems based on the probability of an object. It assumes all the features are conditionally independent but in the case of a real dataset no features are conditionally independent but they can be close.

Linear Regression: Linear Regression is a supervised machine learning approach that is used for predictive analysis. It makes predictions for a continuous or numeric variable.

For all the above cases accuracy values were obtained from the confusion matrix. Table 2 tabulates all the accuracy values. For each run, the error rate was also determined and the time taken to complete the execution was also noted down.

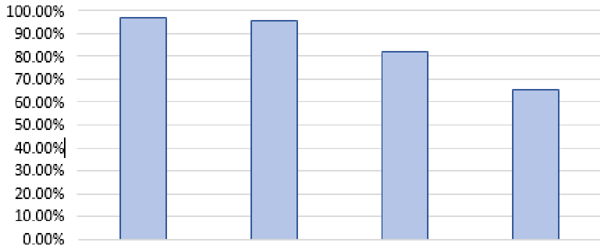


Fig 3. Accuracy of different models using AFINN Analyzer

V. RESULTS AND FINDINGS

A dataset consisting of 1,79,108 tweets was collected for this research. As shown in **Fig 1** and**Fig 2**, the number of neutral, positive, and negative tweets using TextBlob is 111110, 48904, and 19094, respectively, while the number of neutral, positive, and negative tweets using Afinnis 80798, 52433, and 45877.

The sentiment of the tweets was assessed in the first half using two separate analyzers – TextBlob and Afinn. It was discovered that the TextBlob analyzer produces more neutral tweets than Afinn, however after manually analyzing the sentiments of each tweet, it was discovered that Afinn is more precise and accurate, hence TextBlob analyzer has been eliminated for further use.

In the second phase, four algorithms were applied on the Afinn analyzer which is as follows: Logistic Regression produced the highest accuracy of 96.65%, using Decision Tree theaccuracy was 95.51%, by Naïve Bayes the accuracy was 82.09%, and using Linear Regression accuracy was 70.99%.

VI. CONCLUSION

The study of Sentiment Analysis on Twitter Data related to COVID-19 was presented in the research paper. The Prime focus of this research is to find out Positive vs Negative vs Neutral sentiment. It was observed that the highest sentiment from all the tweets, eventuated for Neutral. Almost all countries were expressing their feelings about COVID-19 on various socialmedia platforms, but most of the tweets were obtained from Twitter Web App on #COVID19 for this paper. It had been concluded that Sentiment Analysis using Afinn gives the most accurate result. It had been compared with the manual analysis that Afinn is better than TextBlob. Moreover, Afinn contains 3300+ words with a polarity score associated with each word, and using this library package one can even find the sentiment score of different languages as well.

This work would beneficiate analyzing the sentiments of the people during the pandemic COVID-19 using different types of algorithms. After comparing the results of all the algorithmsthat are Logistic Regression, Decision Tree, Naive Bayes, and Linear Regression, it had been evaluated that the result of Logistic Regression is returning the best result with the highest accuracy. This is because the Logistic Regression algorithm has low variance means less errorin test data so less chance of overfitting. This study provided a good analysis of sentiments andfrom this study, it can be said that the people's reactions vary day to day from posting their feelings on social media specifically on Twitter

VII. FUTURE SCOPES

In the future, the aim is to collect the tweets using Twitter API and plan to create a dataset, as well as use other machine learning algorithms like Hybrid algorithm and then compare the result of the labeled dataset with the result of sentiment analysis that had been performed in this research.

This model can be further taken to new possibilities of Emotion analysis rather than having Positive, Negative, and Neutral tweets. This sentiment analysis model can also be applied to examine the shifting emotions and feelings of individuals and to check if there are noticeable changes over time in them.

REFERENCES

[1] Kausar, Mohammad Abu, Arockiasamy Soosaimanickam, and Mohammad Nasar. "PublicSentiment Analysis on Twitter Data during COVID-19 Outbreak." https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Public+Sentiment+Analys is+on+Twitter+Data+during+COVID19+outbreaks&btnG=&oq=Publ ic+Sentiment+Ana lysis+on+Twitter+Data+during+COVID-19+Outbreak

[2] Abdulaziz, Manal, et al. "Topic based Sentiment Analysis for COVID-19 Tweets." (2021), https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Topic+based+Sentiment+Analysis+for+COVID-19+Tweets&btnG=

[3] Vijay, Tanmay, et al. "Sentiment Analysis on COVID-19 Twitter Data." 2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE). IEEE, 2020, https://ieeexplore.ieee.org/document/9358301

[4] 4. Manguri, Kamaran H., Rebaz N. Ramadhan, and Pshko R. Mohammed Amin. "Twitter sentiment analysis on worldwide COVID-19 outbreaks." Kurdistan Journal of Applied Research (2020): 54-65

[5] Naseem, Usman, et al. "Covidsenti: A large-scale benchmark Twitter data set for COVID-19 sentiment analysis." IEEE Transactions on Computational Social Systems (2021), https://ieeexplore.ieee.org/document/9340540

[6] Khan, Rijwan, et al. "Social media analysis with AI: sentiment analysis techniques for theanalysis of twitter covid-19 data." *J. Critical Rev* 7.9 (2020): 2761-2774, https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Social+media+analysis+w ith+AI%3A+sentiment+analysis+techniques+for+the+analysis+of+tw itter+covid- 19+data&btnG=

[7] Wrycza, Stanisław, and Jacek Maślankowski. "Social media users' opinions on remotework during the COVID-19 pandemic. Thematic and sentiment analysis." *Information Systems Management* 37.4 (2020):288-297, https://www.tandfonline.com/doi/full/10.1080/10580530.2020.18206 31

[8] Pokharel, Bishwo Prakash. "Twitter sentiment analysis during covid-19 outbreak in nepal." *Available at SSRN 3624719(2020)*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3624719

[9] Drias, Habiba H., and Yassine Drias. "Mining Twitter Data on COVID-19 for Sentimentanalysis and frequent patterns Discovery." *medRxiv* (2020), https://www.medrxiv.org/content/10.1101/2020.05.08.20090464v1

[10] Gupta, Prasoon, et al. "Sentiment Analysis of Lockdown in India During COVID-19: A Case Study on Twitter." IEEE Transactions on Computational Social Systems (2020),https://ieeexplore.ieee.org/document/9301194

[11] Pokharel, Bishwo Prakash. "Twitter sentiment analysis during covid-19 outbreak in nepal." Available at SSRN 3624719(2020) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3624719

[12] Raheja, Supriya, and Anjani Asthana. "Sentimental Analysis of Twitter Comments on Covid-19." 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2021, https://ieeexplore.ieee.org/document/9377048

[13] Ghasiya, Piyush, and Koji Okamura. "Investigating COVID-19 News Across Four Nations: A Topic Modeling and Sentiment Analysis Approach." *IEEE Access* 9 (2021): 36645-36656, https://ieeexplore.ieee.org/document/9366469

[14] Machuca, Cristian R., Cristian Gallardo, and Renato M. Toasa. "Twitter Sentiment Analysis on Coronavirus: Machine Learning Approach." *Journal of Physics: ConferenceSeries*. Vol. 1828. No. 1. IOP Publishing, 2021. https://iopscience.iop.org/article/10.1088/1742-6596/1828/1/012104

[15] Hussain, Amir, and Aziz Sheikh. "Opportunities for artificial intelligence–enabled social media analysis of public attitudes toward Covid-19 vaccines." *NEJM Catalyst Innovationsin Care Delivery* 2.1 (2021). https://catalyst.nejm.org/doi/full/10.1056/CAT.20.0649

[16] Shofiya, Carol, and Samina Abidi. "Sentiment Analysis on COVID-19-Related Social Distancing in Canada Using Twitter Data." *International Journal of Environmental Research and Public Health* 18.11 (2021): 5993 , https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8199732

[17] Rehman, Amir, et al. "COVID-19 Detection Empowered with Machine Learning and Deep Learning Techniques: A Systematic Review." Applied Sciences 11.8(2021):3414,https://www.mdpi.com/2076-3417/11/8/3414

LPG Gas Leakage Detector System

Diganta Chakraborty[1], Debajit Jha[1], Sayak Podder[1], Anish Chattopadhyay[1], Koushik Sarkar[2]
Department of Electronics & Communication Engineering,
Future Institute of Engineering & Management
Kolkata -700150

***Abstract*— Since burglaries are increasing daily as a result of the unsafe and unreliable security frameworks in residences, commercial buildings, and enterprises, security may be a significant concern everywhere. We demonstrated a device that can detect the leakage of gas (such as LPG, isobutene, propane, etc.) and alert the customer to the need for action. An alarm that vibrates and sounds like a buzzer is used as a warning device. Buzzer provides an audible indication of the proximity to LPG volume. The LCD and buzzer are turned on by the Arduino UNO. By continuously monitoring residences with various tactile frameworks like vibration, smoke, gas, temperature, door break finders, and fire alarm frameworks, GSM communication frameworks provide security against common, coincidental, expected, unforeseen, accidental, and human-made concerns. The GSM modem continues to deliver SMS messages to mobile numbers that are specifically mentioned in the source code programme to warn people of danger.**

Keywords—component, formatting, style, styling, insert (*key words*)

I. Introduction

These days, gas leakage is observed in many locations, including residences, workplaces, and vehicles like Compressed Natural Gas (CNG), buses, autos, etc. [1-5]. It is noted that risky accidents might occur as a result of gas spills [6]. In many applications, including residences, hotels, companies, cars, and vehicles, liquefied petroleum gas (LPG), also known as propane, is used as fuel due to its attractive qualities, which include a high calorific value, little smoke, little silt, and significant environmental harm [7-9]. After it was discovered that destructive gases had negative effects on human health, placement techniques for gas spills became a worry. Early finding tactics relied on less precise finders until very recently when modern electronic sensors were developed [10].

The number of people dying as a result of gas barrel explosions has been rising for a while now. The Bhopal gas disaster, Mangaluru in Karnataka, Kanpur in Uttar Pradesh, and Valsad in Gujarat are a few of the well-known incidents of gas spillage accidents [11–13].

There have been several audits conducted in the past on the topic of gas spillage location techniques, either as part of research papers or technical studies on specific spill finding strategies and other gas-related topics [14]. An android-based programmable robot for gas finding and signs is shown. The suggested model shows a more adaptable, smaller-than-expected robot capable of detecting gas leaks in hazardous locations [15]. It is said that there are several approaches to identify gas spillages according to the review on gas spill finding and localization processes. They offer a few traditional or contemporary methods to identify the gas.

The processes that are suggested in this work are non-technical, equipment-based tactics that include acoustic, optical, and dynamic strategies [16]. Within the scope of the gas pipeline, spillage point investigation and spillage location investigation have become urgent concerns [17].

One of the most crucial issues nowadays is the extension of the ARM7-based mechanised tall execution framework for LPG refill booking and spillage location and method. It is built on a secretive method that is simple to deconstruct, including an LPG barrel booking unit, a unit for monitoring gas spills at the buyer's end, and a unit for a server framework at the wholesaler's end [18–21]. It has been suggested that using an integrated circuit with MQ-9 might be dangerous for gas detection [22].

To circumvent the problem, Metta Santiputri et al.23 designed a device dubbed the Gas Spill Discovery device based on IoT. (Web of Things). It will continuously monitor the presence of combustible gas in the vicinity, the presence of people, and the proximity of a fire within the building. Another idea for an LPG gas spillage finder based on a microcontroller and a GSM module made use of a gas sensor, GSM module, and microcontroller. The sensors that can detect gas spillage and then communicate with the microcontroller to send signals identify the presence of gas concentration [24].

II. MODELLING OF THE PROJECT

A. Components

- Arduino NANO
- MQ gas Sensor
- GSM Module
- Power Supply
- Wire
- Vero Board
- Buzzer
- Sim card (Other than JIO)

B. Description of the components

Arduino NANO

A coordinated improvement environment built on planning is Arduino. Particularly the implanted framework, physical computing, mechanical technology, computerization, and other hardware-based tasks have been made incredibly simple by it. Each Arduino has about the same functionality and features, with the exception of pin count and measurement. A small chip board based on the AT mega 328p may be the Arduino Nano.

TABLE I. PIN DESCRIPTION

No	Pin Number	Pin Description
1	**D0 – D13**	**Digital Input / Output Pins.**
2	**A0 -A7**	**Analog Input / Output Pins.**
3	**Pin # 3,5,6,9,11**	**Pulse Width Modulation (PWM) Pins**
4	**Pin # 0 (RX), Pin # 1(TX)**	**Serial Communication Pins**
5	**Pin # 10,11,12,13**	**SPI Communication Pins**
6	**Pin # A4, A5**	**12C Communication Pins**
7	**Pin # 13**	**Built-In LED for Testing**
8	**D2 @ D3**	**External Interrupt Pins**

The ATmega328p (Arduino Nano V3.x) / Atmega168-based Arduino Nano is a compact, versatile, and breadboard-friendly microcontroller board created by Arduino.cc in Italy (Arduino Nano V3.x).

Although fairly compact in size, it has precisely the same capability as the Arduino UNO.

It has a 5V operating voltage out of the box, but its input voltage ranges from 7 to 12V.

Each of the 14 digital pins, 8 analogue pins, 2 reset pins, and 6 power pins on the Arduino Nano have many roles, but their primary use is to be setup as an input or output.

When they are connected to sensors, they function as input pins; however, if you are driving a load, you should utilize them as output pins. While analog Read is used to manage the operations of analogue pins, functions like pin Mode and digital Write are used to control the operations of digital pins. The analogue pins measure values between 0 and 5V with a total resolution of 10 bits. A crystal oscillator with a 16 MHz frequency is included with the Arduino Nano. It is used to generate an accurate clock with a steady voltage. One drawback of utilizing an Arduino Nano is that it lacks a DC power connection, which prevents you from using a battery as an external power source.

Instead of using regular USB to connect to a computer, this board has support for Mini USB. This device is a superb option for the majority of applications where the sizes of the electrical components are a major consideration due to its small size and breadboard-friendliness.

Depending on the At mega board, flash memory can be 16KB or 32KB. For example, the Atmega168 has a 16KB flash memory whereas the Atmega328 has a 32KB flash memory. To store code, utilise flash memory. Out of the entire flash memory, a bootloader uses 2KB of memory.

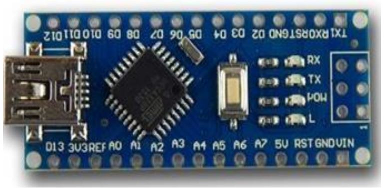


Figure 1: Arduino NANO

For the Atmega168 and Atmega328, the SRAM can range from 512 bytes to 2KB, while the EEPROM can be either 512 or 1KB. Although this board is relatively comparable to other Arduino boards on the market, its compact size sets it apart from the competition. The specs for the Arduino Nano Board are shown in the following figure. It is programmed with the Arduino IDE, an offline and online integrated development environment. To operate the board, no special preparations are necessary. Board, a tiny USB cable, and Arduino IDE software already installed on a computer are all you need. The programme is sent from the computer to the board via a USB connection. No separate burner is required to compile and burn the program as this board comes with a built-in boot-loader.

TABLE II. PIN DESCRIPTION

Microcontroller	Atmega328p/Atmega 168
Operating Voltage	5V
Input Voltage	7 – 12V
Digital I/O Pins	14
PWM	6 out of 14 digital pins
Max. Current Rating	40mA
USB	Mini
Analog Pins	8
Flash Memory	16KB or 32KB
SRAM	1KB or 2KB
Crystal Oscillator	16 MHz
EEPROM	512bytes or 1KB
USART	yes

Following figure shows the pinout of Arduino Nano Board.

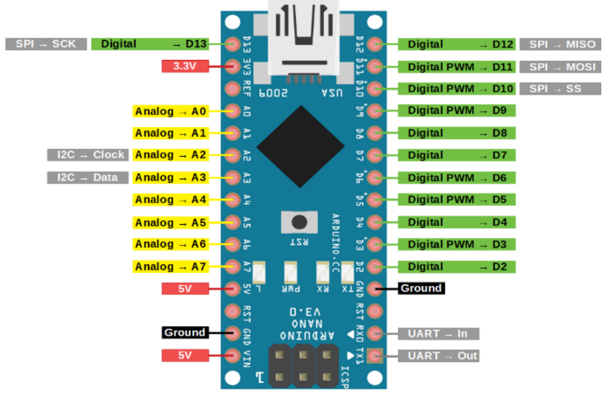


Figure 2: Arduino NANO Pin Description

Every pin on the Nano board has a specific job associated with it. Are you ready to view the analogue pins that may be used as an analogue to digital converter? You can also use the A4 and A5 pins for I2C communication. There are really 14 computer pins total, of which 6 are used to generate PWM.

PIN Description

V_{in}: It is input power supply voltage to the board when using an external power of 7 to 12 V.

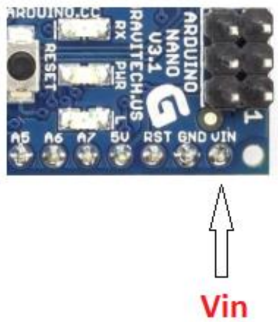


Figure 3: Vin pin.

5V: it is a regulated power supply voltage of the board that is used to power the controller and other components placed on the board.

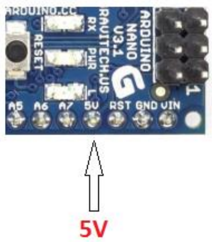


Figure 4: 5V pin.

3.3V: This is a minimum voltage generated by the voltage regulator on the board.

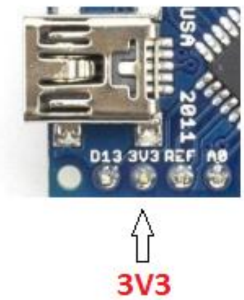


Figure 5: 3V3 pin..

GND: Multiple ground pins are on the board that can be interfaced accordingly when more than one ground pin is required.

Figure 6. Ground Pin

Reset: Every pin on the Nano board has a specific job associated with it. Are you ready to view the analogue pins that may be used as an analogue to digital converter? You can also use the A4 and A5 pins for I2C communication. There are really 14 computer pins total, of which 6 are used to generate PWM.

Figure 7: Reset

Analog Pins: There are 8 analog pins on the board marked as A0 – A7. These pins are used to measure the analog voltage ranging between 0 to 5V.

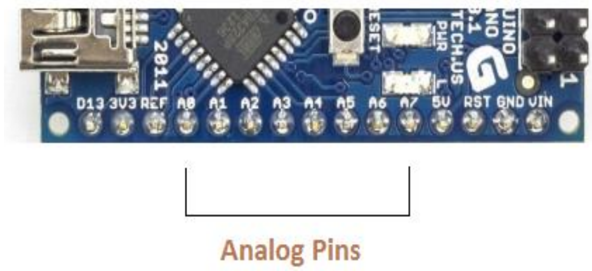


Figure 8: Analog pins

R_x, T_x: These pins are used for serial communication where T_x represents the transmission of data while R_x represents the data receiver.

Figure 9: Rx and Tx pin.

13: This pin is used to turn on the built-in LED.

AREF: This pin is used as a reference voltage for the input voltage.

PWM: Six pins 3, 5, 6, 9, 10, 11 can be used for providing 8-pit PWM (Pulse Width Modulation) output. It is a method used for getting analog results with digital sources.

SPI uses four pins: 10 (SS), 11 (MOSI), 12 (MISO), and 13 (SCK) (Serial Peripheral Interface). Data is typically sent between microcontrollers and other peripherals like sensors, registers, and SD cards via the interface bus known as SPI.

External Interrupts: Pins 2 and 3 are used as external impediments, which are used in times of emergency when we need to pause the majority of our programmes and ask for urgent education at that time. Once a hinder instruction is called and carried out, most programmes restart.

I2C: I2C communication is made possible by the A4 and A5 pins, where A4 communicates with the serial information line (SDA), which carries the information, and A5 communicates with the serial clock line (SCL), which could be a clock flag created by the master device and used to synchronise the information between the devices on an I2C transport.

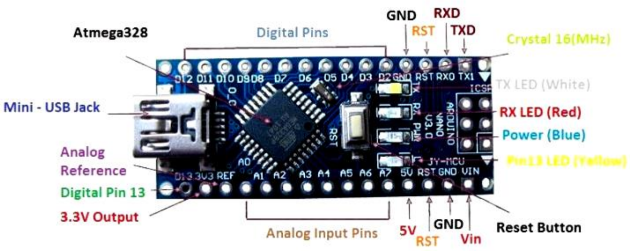


Figure 10: Arduino NANO.

Communication and Programming

A communication channel may be established between the Nano device and other controllers and computers. The more sophisticated pins, such as stick (Rx) and stick 1 (Tx), are used for serial communication, with Rx being used for information reception and Tx being used for information transfer. The Arduino computer application has a serial screen that may be used to send or receive printed data to or from the board. Additionally included in the application are FTDI drivers, which serve as a virtual com port for the computer programme. As data is sent between an FTDI and USB connection to the computer, the Tx and Rx pins have a Driven that flickers.

For carrying out a serial communication between the board and the computer, the Arduino Program Serial Library is used. The Nano sheets provide I2C and SPI communication apart from serial communication. The Arduino program's Wire Library is accessed in order to use the I2C transport. The Arduino computer application called IDE, which is frequently used for almost all types of boards available, modifies the Arduino Nano. Download the computer application, then choose the board you want to use.

There are two ways to programme the controller: either using the bootloader built within the computer software, which frees you from needing an external burner to build and burn the programme into the controller, or via ICSP (In-circuit serial programming header). Although the Arduino board software is compatible with Linux, Mac, and Windows, Windows is the most common.

GSM Module

This GSM modem functions just like a mobile phone with its own unique number and can take any GSM network operator SIM card. The RS232 connector, which may be used for communication and the creation of embedded programmes, is a benefit of utilising this modem. It is simple to create applications for SMS control, data transmission, remote control, and logging.

The modem may be connected to any microcontroller or especially to the serial port on a PC. It may be used to place and receive voice calls as well as send and receive SMS. Additionally, it may be used in GPRS mode to connect to the web and perform a variety of data recording and control applications. Additionally, you can connect to any unreachable FTP server in GPRS mode and transfer files for information logging.

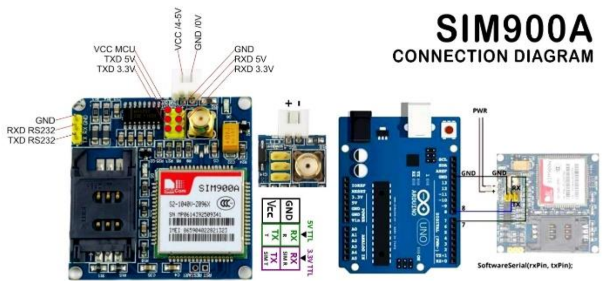


Figure 11: GSM Module.

MQ Gas Sensor

These days, almost every family uses LPG for cooking. LPG is widely used in industrial and commercial setups in addition to being used for domestic applications. Utilizing LPG Gas is unquestionably very beneficial, but if the correct precautions aren't taken, it can cause severe accidents, damage to property, and even perhaps loss of life.

We may avoid the harm caused by any real situation by using one essential and simple device. LPG Gas Sensors are tools that can foresee any unnecessary injury or an accident. These days, the industry offers a wide variety of LPG Gas Sensors. Depending on the type of hardware it has to be installed on, you may choose the one that best meets your needs. It is crucial to take into account the gas detector's height as well as its width.

Figure 12: Nair, N. (n.d.). MQ-5 LPG GAS SENSOR. ElementzOnline. Retrieved June 9, 2022, from https://www.elementzonline.com/mq-5-lpg-gas-sensor-289.

When purchasing an LPG gas sensor, it's important to keep in mind that it should be extremely sensitive to the operator doing the noticing within the LPG barrels. As a result, it would be able to recognise the fragrance immediately and permit a flag. The gas sensors should be designed such that they can withstand the aroma of alcohol, smoke, or other fragrant materials. This sensitivity to LPG is crucial in order to prepare for false alarms when it detects the scent of anything other than LPG.

The discovery range of an LPG gas sensor is another significant feature. The LPG gas sensor should be sensitive enough to detect the faintest gas aroma. In addition to that, the LPG gas sensor's response time should be taken into account. A sensor with a quick reaction time can identify an LPG spill right away. In this manner, sensors with a slow reaction time appear to have the best results and immediately send out warning signals, therefore predicting any mishap. Usually an easy-to-use LPG Gas Sensor Module that can detect the presence of flammable gases such as LPG, isobutene, and propane in the vicinity. The sensor used by the module is the MQ-5.

It simplifies interaction to the sensor's odd stick division and provides interface through four 0.1" header pins. It provides a computerised yield that is easy to use as well as an analogue yield that is in accordance with the gas concentration under discussion. The highest gas concentration past which the computerised yield becomes triggered may be adjusted using the onboard potentiometer.

Fairly controlling the module with 5V, setting the edge, and hopefully yielding results! An onboard Driven alerts the crew to the presence of any gas. The improved yield easily interacts with other circuits, including microcontrollers. To enable a broad range of sensor trawling, the analogue output may be snared up to an ADC of a microcontroller.

5 Volt Power Supply

A three terminal positive voltage controller with a 5V settled potential is the 7805. The IC is very durable since it includes features like internal current restricting, warm shutdown, and secure functioning range guarantee. Given that there is an enough warm sink, yield streams of up to 1A can be extracted from the IC. The most voltage is reduced by a 9V transformer, which is then amended by a 1A centre tap, channelled by capacitor C1, and controlled by a 7805. The result is a steady 5Volt DC. The circuit diagram is provided below.

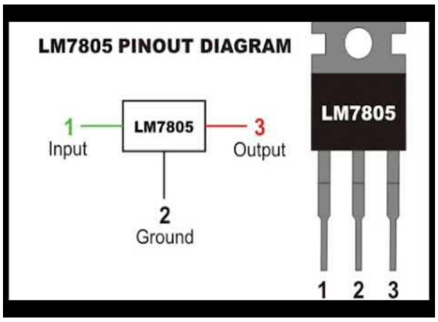


Figure 13: LM7805 Pinout diagram. (2020, December 22).LPG gas leakage detector. Retrieved June 9, 2022, from https://www.ecstuff4u.com/2019/09/lm7805-pinout.html.

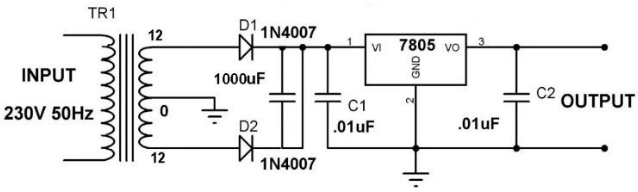


Figure14: Input and Output.

Block Diagram

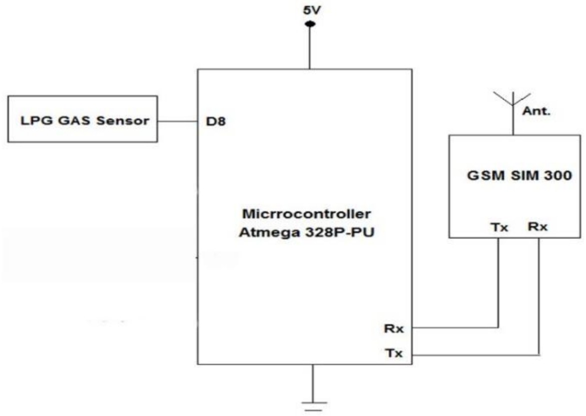


Figure 15: Block Diagram.

Results and discussions

Each sensor circuit's output is sent to the corresponding mono-shot trigger circuit for comparison. The Arduino NANO's input stick receives the output from the trigger circuit, which prompts the GSM SIM 300 modem to send an SMS to the specified number. As a consequence, we will see that the buzzer will start to tune when any gas is detected in the surroundings. The sensor will communicate the yield to the microcontroller, which is then able to send the yield to the designated phone number by way of sms using the GSM module.

The figure 16 shows the prototype of the proposed system.

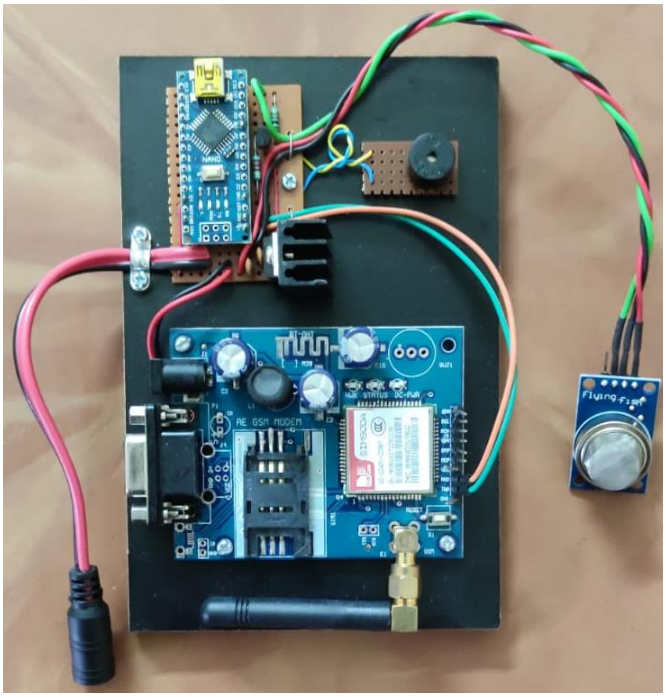


Figure 16: Prototype of our own system

FUTURE SCOPE

A variety of circuits may be combined on a single Arduino pack, giving it a broad range of expansion. Since a GSM pack with a SIM card is used to facilitate communication, the operation's scope may expand in the near future. The inclusion of a sub framework where gas wastage may be detected using this framework is one of the outstanding future tasks for this framework. Usually, a preprogramed framework for gas discovery, control, and alarm. In the future, this framework will include a feature where it may alert the crisis administrations if an accident occurs recently.

REFERENCES

[1] Mahalingam, A.; Naayagi, R.T.; Mastorakis, N.E. Design and implementation of an economic gas leakage detector. In Proceedings of 6th International Conference on Circuits, Systems and Signals, Athens, Greece, 7-9 March 2012; pp. 20-24.

[2] Attia, H.A.; Halah, Y.A. Electronic Design of Liquefied Petroleum Gas Leakage Monitoring, Alarm, and Protection System Based on Discrete Components. Int. J. Appl. Eng. Res. 2016, 11, 9721-9726.

[3] Apeh, S.T.; Erameh, K.B.; Iruansi, U. Design and Development of Kitchen Gas Leakage Detection and Automatic Gas Shut off System. J. Emerg. Trends Eng. Appl. Sci. 2014, 5, 222-228.

[4] Soundarya, T.; Anchitaalagammai, J.V.; Priya, G.D.; Karthickkumar, S.S. C-Leakage:Cylinder LPG Gas Leakage Detection for Home Safety. IOSR J. Electron. Commun. Eng. 2014, 9, 53-58.

[5] Shrivastava, A.; Prabhaker, R.; Kumar, R.; Verma, R. GSM based gas leakage detection system. Int. J. Emerg. Trends Electr. Electron. 2013, 3, 42-45.

[6] Anurupa, A.; Gunasegaram, M.; Amsaveni, M. Efficient Gas Leakage Detection and Control System using GSM Module. Int. J. Eng. Res. Technol.2015,3, 1-4.

[7] Meenakshi, A.A.; Meghana, R.B.N.; Krishna, P.R. LPG Gas Leakage Detection and Prevention System. Int. J. Future Revolut. Comput. Sci. Commun. Eng. 2017, 3, 1-4.

[8] All Answers Ltd. GSM Based LPG Detection [Internet]. November 2018. Available online: https://ukdiss.com/examples/gsm-based-lpg-detection.php?vref=1 (accessed on 15 October 2020).

[9] Taufiq Noor Machmuda, LPG Gas Detector and leak prevention based microcontroller, 2009.

[10] Mahalingam, A., R. T. Naayagi, and N. E. Mastorakis. "Design and implementation of an economic gas leakage detector." Recent Researches in Applications of Electrical and Computer Engineering, pp. 20-24, 2012.

[11] Attia, Hussain A., and Halah Y. Ali. "Electronic Design of Liquefied Petroleum Gas Leakage Monitoring, Alarm, and Protection System Based on Discrete Components." International Journal of Applied Engineering Research, vol. 11, no. 19, pp. 9721-9726, 2016.

[12] Apeh, S. T., K. B. Erameh, and U. Iruansi. "Design and Development of Kitchen Gas Leakage Detection and Automatic Gas Shut off System." Journal of Emerging Trends in Engineering and Applied Sciences, vol. 5, no. 3, pp. 222-228, 2014.

[13] T.Soundarya, J.V. Anchitaalagammai, G. Deepa Priya, S.S. Karthick kumar, “C-Leakage: Cylinder LPG Gas Leakage Detection for Home Safety,” IOSR Journal of Electronics and Communication Engineering, vol. 9, no. 1, Ver. VI, pp. 53-58, Feb. 2014.

[14] Ashish Shrivastava, Ratnesh Prabhaker, Rajeev Kumar, Rahul Verma, "GSM based gas leakage detection system." International Journal of Emerging Trends in Electrical and Electronics, vol. 3, no. 2, pp. 42-45, 2013.

[15] Shrivastava, A., Prabhaker, R., Kumar, R., & Verma, R. GSM based gas leakage detection system. International Journal of Emerging Trends in Electrical and Electronics (IJETEE-ISSN: 2320-9569), 2013; 3(2):42-45.

[16] Hema, L. K., Murugan, D., & Chitra, M. WSN based Smart system for detection of LPG and Combustible gases. In National Conf. on Architecture, Software systems and Green computing-2013.

[17] Ramya, V., & Palaniappan, B. Embedded system for Hazardous Gas detection and Alerting. International Journal of Distributed and Parallel Systems (IJDPS), 2012; 3(3):287-300.

[18] Priya, P. D., & Rao, C. T. Hazardous Gas Pipeline Leakage Detection Based on Wireless Technology. International Journal of Professional Engineering Studies, India, 2014; 2(1).

[19] Jero, S. E., & Ganesh, A. B. 2011, March. PIC18LF4620 based customizable wireless sensor node to detect hazardous gas pipeline leakage. In 2011 International Conference on Emerging Trends in Electrical and Computer Technology (pp. 563-566). IEEE.

[20] Anusha, O., & Rajendra prasad, C. H. Experimental investigation on road safety system at crossings. International Journal of Engineering and Advanced Technology, 2019; 8(2):214–218.

[21] Pravalika, V., & Rajendra Prasad, C. Internet of things based home monitoring and device control using Esp32. International Journal of Recent Technology and Engineering, 2019; 8(1 Special Issue 4):58–62.

[22] Sanjay Kumar, S., Ramchandar Rao, P., & Rajendra Prasad, C. Internet of things based pollution tracking and alerting system. International Journal of Innovative Technology and Exploring Engineering, 2019; 8(8):2242–2245.

[23] Deepak, N., Rajendra Prasad, C., & Sanjay Kumar, S. Patient health monitoring using IOT. International Journal of Innovative Technology and Exploring Engineering, 2018; 8(2):454–457. https://doi.org/10.4018/978-1-5225-8021-8.ch002.

[24] Ramu, M., & Prasad, C. R. Cost effective atomization of Indian agricultural system using 8051 microcontrollers. International journal of advanced research in computer and communication engineering, 2013; 2(7):2563-2566.

Different Approaches for Multi-Class Classification using Machine Learning Techniques

Prashant Y Niranjan, Vijay S Rajpurohit

Department of Computer Science & Engineering, KLS Gogte Institute of Technology, Belagavi.

ABSTRACT

Web content increasing in every sector has become challenging task to find the useful information. Question Answering system in agriculture domain help farmers to provide the accurate answer. Farmer asking the queries relevant to pomegranate fruit in which question classification plays an important role for various questions asked by farmer to categorize or to identify the type of question. Different question classification methods have been proposed to provide solutions for classification of question. In this research, we are considering the different pomegranate questions that can be asked by farmer to classify the questions correctly using different machine learning methods. A proposed framework for question classification having multiple classes i.e. name, descriptive, location, numeric and entity: other which enables machine learning algorithms to categorize the type of question. This paper compares various machine learning algorithms and result shows K-Nearest Neighbor, SVM & Decision tree performed well with good accuracy.

Keywords: Question classification; Machine learning, Natural Language Processing, Text Mining.

1. Introduction

Agriculture is the need of every country and it increases the economy of a country. Providing the needs of our farmer is one of the important tasks to carry out their work effectively and smoothly. In this research pomegranate fruit has been considered, Pomegranate is a commercial fruit plant which belonging to Punicaceae family [1]. It has good source of protein, minerals, antioxidants, carbohydrate, vitamins A, B and C, also it's used for restrain or to control heart diseases, cancer, fever, leprosy, abdominal pain etc. Pomegranate is in more demand; its production is increasing day by day and exporting over the past years and earning higher profit from the production.

Huge amount of data flow becomes very difficult to get the accurate content which the user is looking for. Current search engine do not have eliminating capability to reduce the content and providing the exact answer to users. Normally it gives number of URL for the user asked questions which makes tedious and time consuming task for the user to open the correct link to find the required or exact answer.

Considering these drawbacks we have proposed the development of question answering system. Farmer growing the pomegranate in his field has lot of queries about various parameters which may involve water, growing methods, soil, disease etc., hence QA system has been introduced to provide solution to every user according to their requirement and can be used in all sectors.

Question answering system is used to provide the intercommunication between human languages and computer, it programs the machine to understand and analyze the large amounts of human language data. QA is a slice of NLP, wherein developing the farmer Question answering system that has been trained to provide automatic responses to the questions asked by farmer in natural language to receive the exact answers.

QA system has become one of the booming areas of research, for identification of questions types question classification module plays a major role in the classification process.

QA system has become user friendly, improving the user needs and methods/techniques used in finding the associated results, because of drastic growth in the amount of web content and getting lesser responses to the asked queries. For the expected answer type, the task of classification process is to assign the accurate class outcomes to the queries asked by the user.

Based on class outcome of a question, it displays the reply to the user questions. So performing classification of questions directly affects the results or answers, miss-classification of questions in QA system leads to the erroneousresults.

Now a day most of the farmers are literate, they have good ideas about how to generate the revenue from their growing field, based on the climate, soil and water conditions most of the people choose some items like fruits, vegetables or some grains to grow in their field. To grow any kind of item we need support of technology to get some help in online mode and the same kind of techniques/tools/procedures can be applied by the farmer to grow in his/her field.

Farmer growing the pomegranate has number of queries like water, soil, new techniques/technologies/tools and disease related issues. He/she typing the question for his problems and the type of question will be classified as Descriptive, Name, Location, Numeric & Entity: other, this will help our system to extract the exact answer to the farmer. Here the task of proving accurate answers to farmers is more dependent on the question type; hence question classification task directly affects the answer. Based on expected answer type, classification task is used to allot the appropriate class outcome to questions.

Most used algorithm for multi-class classification is Support Vector Machine & Logistic regression; even other ML algorithms used are Nearest Neighbor, Decision Tree, Naive Bayes & Neural networks.

In this research, we propose the pomegranate data set consisting of 300 questions built from authentic NRCP file [1]. Here comparison of different ML algorithms for question classification process has been performed for the asked questions.

This paper is organized as: Introduction of the paper is presented in Section 1; Previous work carried out in QA system and machine learning techniques used for question classification approaches outlined in section 2. Section 3 discusses the research objectives. Different question classification approaches are described in section 4. The experimental setup and results are shown in Section 5; results are discussed in Section 6. Finally, Section 7 concludes the paper and outlines directions for future work.

2. Background

Previous work on question classification based on user intent has been discussed in this section. Different class labels of user intent are described in Section 2.1, previous work on question classification techniques/methods are described in section 2.2., and user asked questions the figure 1. shows the different steps applied for QA system to get the accurate answer. Different algorithms or techniques have been provided for the different modules of QA system.

A. Question Categories

Different class labels for questions are defined, which is shown in below table 1. Major question types are: Descriptive type, Name, Numeric, Location and Entity: other.

Table 1. Categories of question.

Authors	Categories
[2]	definition, casual, relationship, Factoids, list, procedural, hypothetical, and confirmation questions
[3]	Solution, Definition, Navigation, Fact, List, Reason,
[4]	Cause and effect, Definition, Rationale, significance Advantage/Disadvantage, Comparison, Explanation, Identification, List, Opinion,
[5]	Entity. Human, Location and Numeric as coarse classes; colour, city, Abbreviation, Description

Different question labels are available like factoid, definition, procedural, List, description, opinion, solution, explanation & Reason all these classes are available under the Descriptive type question. A descriptive type question usually starts with Wh-type questions where it normally uses what and How keyword.

For Name type label, it is used to find the answer for name type question. It uses the WHICH keyword to ask the questions and needs to find the name that can be fruit name, vitamin name, disease name or seed name etc.

For Numeric label, it requires to find the numeric relevant answers like quantity, value, measurement, feet, inch, weight and profit/loss. It uses how much, What, When, How many keywords are used to ask the questions to find the numerical related answer to the user.

For Location label, it requires to find the location relevant answers like place, country, state and area. It uses WHERE keyword to ask the questions to find the location related answer to the user.

For Entity: other label, it requires to find the kind of object information except from the given labels. User is asking the question relevant to any real time entity or other questions will be classified as Entity: other to find the entity related or any other answer to user.

2.2 Question classification methods

This section discusses the previous work carried out on question classification methods and used different ML algorithms.

Authors in Alaa Mohaseeb et al [7]. have proposed using grammatical features for Classification of factoid questions intent, they have used techniques of SVM and J48 and it contains the Grammatical features the word like what, which, who, when. Domain specific features contain the features like religious terms, health terms, and events. Grammatical patterns have structure for type of question. It has been proposed for type of question to be classified and have limitation where rule based and grammatical patterns are used and it is not accessing the recent and authorized links.

Muhammad Wasim et al. proposed the Classification of Factoid and List type Questions in Biomedical Question Answering for Multi-Label Question [8]. Authors have highlighted the *technique of binary* relevance transformation method and with consideration of extracting features (Question Focus). It's been used to classify list and factoid type questions but using this technique it's a time consuming process.

Bastian Haarmann et al. have proposed the [9] Mighty Data set for stress testing QA system, which uses SPARQL Queries. It's used to extract data from fixed set of NLP questions. Its limitations is that it requires separate SPARQL queries to be written for all different types of questions and works only for fixed set of questions.

Hasangi Kahaduwa et al [10] proposed the Question Answering system for the travel domain. Techniques used are linear SVM algorithm, rule based approach and SPARQL queries are used. It has been designed to provide the answer for travel related queries and works only for fixed set of NLP questions.

M.R. Sumalatha et al [11] "*Analyzing and predicting knowledge of contributors in community question answering services" Here the* knowledge contributors are involved and its based on the number of likes and votes. This works good for all the domain related queries and its open for all the people question and answers. This system is very time consuming and it cannot assure that all responded answers are accurate.

Sharvari Gaikwad et al [12] ***"AGRI-QAS question-answering system for agriculture domain" [12]*** have used the NER and POS tagging , system works only for factoid questions such as 'which', 'what', 'where', 'who' and for semantic type questions it does not perform well.

Payal Biswas et al [13] "A framework for restricted domain Question Answering System" have used the Alchemy content extractor, Paragraph extractor, standford coreNLP toolkit, WH-type questions, head word. Here answer extraction works good when the keywords, Headword is present and the answer statement is in correct format. Here rule based approach is used, if answer statement is not in the proper format answer extraction fails to work.

Haarmann et al. (2018) have proposed [14] the QA system which has mighty dataset of stress-testing and it contains large number of NLP questions and SPARQL queries are used.NLP questions are in fixed format and it uses SPARQL queries. Here a given natural language question is converted into a suitable SPARQL query that displays the correct answer.

Gautre et al. (2018) have proposed [15] a QA system which has the submitted question and produces it to the speaker, crowd sourcing application designed to produce the suitable answer to the questioner.

Devi and Dua (2017) et al. [16] developed a QA system for agriculture domain using ontology. ontology web language and Resource description form of data available. SPARQL protocol and RDF query language is used to extract this data.

3. Research Objective

In order to provide valuable support to the farmers in decision making by providing the precise answers to the farmers when they don't have knowledge about the asked query.

In this system firstly it finds out the type of question, expected type of answer, getting keywords of the question and focus of the question then it collects or retrieves the data from the WWW, news, articles etc and then finally it extracts the exact answer using the trained model which is developed using machine learning techniques.

QA system it performs the following functions:

i) Question processing module
ii) Document processing module
iii) Answer extraction module

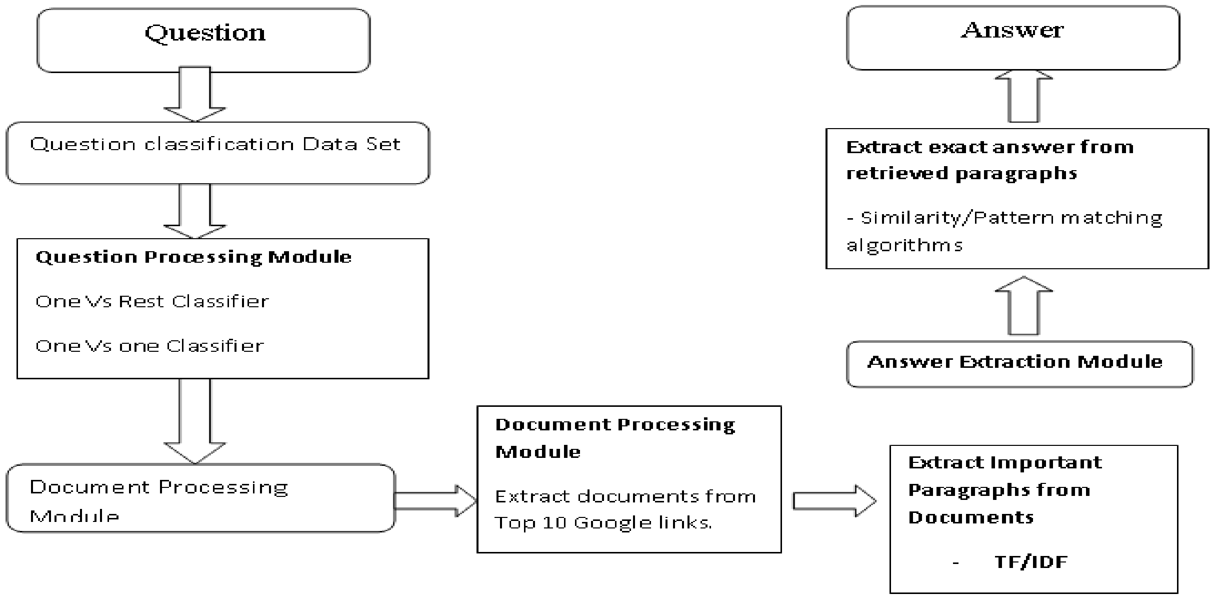


Fig. 1 Question Answering System Architecture

Our research is on QA system for particular or closed domain. Mainly the QA system is used for providing the accurate answer to the user and it has mainly three steps to provide the exact answer. In this paper we are working on question classification task to classify the given question accurately for the expected answer type.

Our research focuses on QA system for pomegranate fruit, Farmers those who are into agriculture and growing specifically the pomegranate fruit has number of queries, especially those are new into farming of this fruit. To provide solution for this problem we are designing and developing a QA system which helps all the famers.

In this paper, we are working on question classification task of pomegranate QA system which is used to provide the expected answer type according to the categorization of question.

The goal of this research paper is to:

1. Investigating the different types of pomegranate relevant questions under multi-class to check the performance ofquestion classification system.
2. Evaluating and comparing the different ML algorithms used for classification task of a question to check theperformance of different algorithms.
3. Evaluating the result of each algorithm to conclude the best algorithm of machine learning can be used for multiclass classification.

In the first objective some sample question figure/questions will be taken and illustrations of data set with different categories are discussed.

Second Objective is to compare and discuss the various ML algorithms for classification of question.

Third objective shows the result of each algorithm and discussing the same with code implementation. Afterdiscussing the result part of each algorithm and it concludes with the highest accuracy giving algorithm.

4. System Architecture

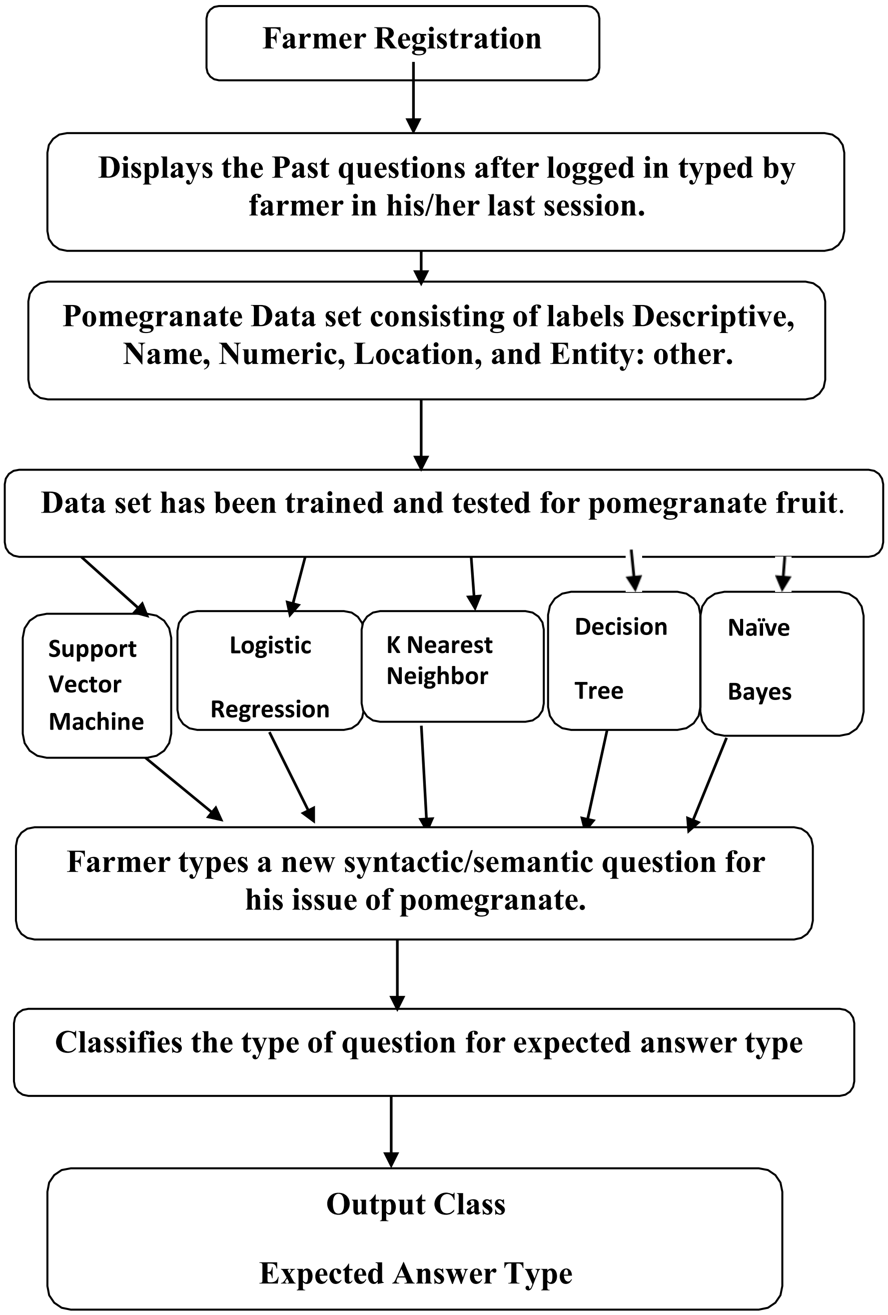


Fig. 2 System Architecture of Question Classification

4.1. Farmer Registration

Farmer growing the pomegranate has number of queries in his mind to ask the question to QA system. First the farmer needs to register to QA system with his details and mobile number can be taken as key for registration. Next time when he logs in he can his username as mobile number and the entered password. Once he login into QA system he can ask any kind of queries relevant to pomegranate.

4.2. Displays the Past questions after logged in

It displays the past questions which are entered by the user. After registration user has entered many questions in QA system, but after every login it displays the past or some previous questions entered by the user.

4.3. Pomegranate Data set consisting of classification labels

Pomegranate data set consisting of multi class labels specified for the number of questions to produce the output. It has labels like Descriptive, Name, Location, Numeric and Entity: other labels.

Class Label	Answer
Descriptive	Provides two or more lines of answer, definition types of answer.
Name	Provides the name of the fruit, disease, seed, pesticide etc.
Numeric	Provides quantity, pounds, rupees, time and date etc.
Location	Provides city, state, country and area.
Entity: Other	Provides other types of answer.

4.4. Data set has been trained and tested of pomegranate fruit

Pomegranate data set have many or number of questions considering the several parameters like its growth, disease, water, pesticides etc. Here some questions of data set have been trained and some questions of data set have been included for testing.

Numbers of classification algorithms have been used like logistic regression, K-Nearest Neighbor, Decision Tree, support vector machine, and Naïve Bayes algorithm.

4.5 Farmer types a new syntactic/semantic question.

Farmer who grows the pomegranate has number of issues, once the data has been trained and tested or model has been built using number of classification algorithms, farmer asking the query relevant to pomegranate for his issue. Farmer has number of queries in his mind while growing the pomegranate specially when his new into farming or into growing of pomegranate fruit.

4.5 Classifies the type of question for expected answer type

Farmer typed question it classifies the type of question to produce the expected answer type, it involves the number of labels/category like Descriptive type, Name, Location, Numeric and Entity:other.

4.6 Expected Answer Type

Finally it displays the result that is expected answer type to the farmer.

5. Experimental Study and Results

This section examines the different machine learning classifiers to categorize on different question types. Pomegranate data set for question classification is important component to have one of the predefined categories for the farmer entered question. Data set for question classification task has been prepared by using NRCP pomegranate file [1]. For our research work 300 pomegranate questions has been prepared along with outcome labels that is Descriptive, Name, Numeric, Location and Entity: Other.

Slno	Question	Classlabel
1	what are the ailment symptoms of bacterial leaf spot?	DESC:desc
2	What are the sickness signs and symptoms of baterial fuit spot?	DESC:desc
3	what is the predominant source of inoculum in contaminated cuttings?	ENTITY:other
4	what is the optimum temperature for the initiaion of fruit spot in pomegranate?	Numeric
5	what is the beneficial climate situation for the inititation of leaf spot in pomegranate?	DESC:desc
6	what is the favourable temperature for the initiation of leaf spot in pomegranate?	Numeric
7	which disease is the most frequent in pomegranate?	Name
8	what are the disease signs of anthracnose?	DESC:desc
9	what is the source of foremost spread of leaf spot in pomegranate?	DESC:desc
10	how can i forestall leaf spot from spreading?	DESC:desc
11	how can i forestall the fruit spot from spreading?	DESC:desc
12	what are the favourable prerequisites for anthracnose?	DESC:desc
13	what are the sickness symptoms of Fusarium wilt?	DESC:desc
14	what are the signs and symptoms of Alternaria fruit spot?	DESC:desc
15	what are the beneficial weather condition for bacterial blight?	DESC:desc
16	what are the reasons for bacterial blight?	DESC:desc
17	what is the beneficial weather for the unfold of anthrecnose?	DESC:desc
18	what are the signs of cercospora fruit spot?	DESC:desc
19	what are the signs of cercospora leaf spot?	DESC:desc
20	What are the signs and symptoms of stem-canker?	DESC:desc

Fig. 4 Pomegranate Questions with Output Labels (Supervised Data set).

To convert your words into numbers. To process machine learning algorithm on a sentence. These words cannot interpret by machine learning algorithm. So we need to convert these words to numbers.

- ► Word2Vec
- ► Count Vectorizer
- ► TF-IDF Vectorizer

Count Vectorizer

Text= ["hello my name is aman and I am a data scientist"]

Text1=["you are watching unfold data science aman"]

0	1	2	3	4	5	6	7	8
Am	Aman	And	Data	Hello	Is	my	Name	scientist

Fig 5 Representation of word embedding.

After preprocessing the data in background the above diagram shows the representation of word embedding. With reference to text1 below is the output

[0 1 0 1 0 0 0 0 0]

If there are common words between these it will be 1. This will not be helpful when documents or domain is different. If text1 is about cricket and text2 is about football. This will not be helpful. So we have TF-IDF vectorizer.

TF-IDF Vectorizer

Text = [" Aman is data scientist in india" , This is unfold data science", "Data science is promising carrer"]Above text has 3 documents, After preprocessing the data word to vec is created.

0	1	2	3	4	5	6	7	8	9	10
aman	carrer	Data	In	India	Is	Promising	science	scientist	This	unfold

Fig 6 Representation of word embedding

Word vec will be compared with the given input sentence if its present it will be 1 otherwise it will be 0.Text as input = text[0]

Text [0] = "aman is data scientist in India"[1 , 0 , 1, 1, 1, 1]

Text [2] = "Data science is promising career"[0 ,1 , 1, 0, 0 , 1 , 1 , 1]

Different machine learning algorithms have been used for question classification. These are briefly described below. Multi-class classification is not supported by all classification models; for binary classification designed algorithms such as the Perceptron, Logistic Regression, and Support Vector Machines and for more than two classes these algorithms do not support.

Multi-classification problems can be solved by using binary classification algorithms which splits the multi-class classification dataset into multiple binary classification datasets to fit a binary classification model on each.

Binary Classification: For two classes Classification.

Multi-class Classification: For more than two classes Classification.

For Binary classification problems some of the designed algorithms are which includes: Support Vector Machines, Logistic Regression, Perceptron, but they cannot be used for multi-class classification tasks that too not directly.

Two different approaches used are

► One Vs Rest and

► One Vs One

5.1 Logistic Regression

For multi-class classification, logistic regression uses the technique of One-vs-rest which uses binary classification algorithms that splits the multi-class dataset into multiple binary classification problems. For each binary classification problem, a binary classifier is trained and predictions are made using the model that is the most confident. It uses one-versus-rest approach, in which we train C binary classifiers, fc(x), where the data of class c is treated as positive, and data from all the other classes is treated as negative.

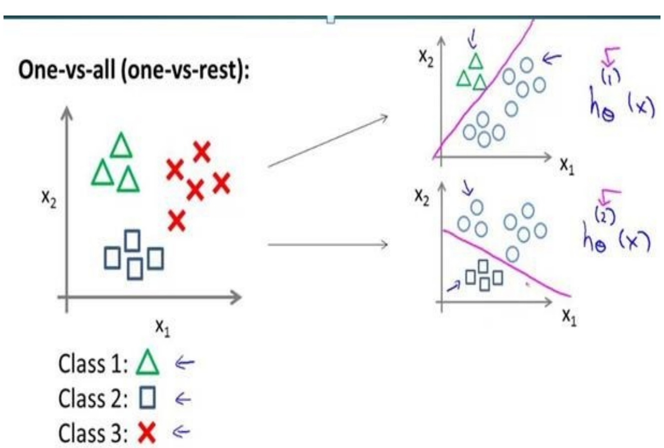


Fig.7 One Vs Rest Classifier

For 3 categories one group will be positive and other 2 group will be negative (M1 model is created), For M2 model other side one group will be positive and other group will be negative and continues.

For M1 model it gives one probability value 0.20, M2 model gives probability value 0.25 and M3 model gives 0.55. Then it will check which has given the highest probability value, M3 model has given highest probability value so it means given new data belongs to O3 category.

Logistic regression algorithm has been applied on the supervised data set and for the given new data; it's been classified into the appropriate category. For 300 questions supervised data set has been applied with logistic regression and it is giving the accuracy of 68.85%. For K-fold cross validation the test accuracy score mean is 57.38% and it shows the below confusion matrix.

Table. 1 Example of One Vs Rest Classifier

f1	f2	f3	O/P	O1	O2	O3
I1	I2	I3	O1	+1	-1	-1
I4	I5	I6	O2	-1	+1	-1
I7	I8	I9	O3	-1	-1	+1
I10	I11	I12	O1	+1	-1	-1
I13	I14	I15	O2	-1	+1	-1
I16	I17	I18	O3	-1	-1	+1

```
#mythreshold=0.5
from sklearn.metrics import confusion_matrix
#predicted= text_clf.predict(X_test)>= mythreshold).astype(int)
cm = confusion_matrix(y_test, predicted)
print(cm)

print(metrics.classification_report(y_test, predicted, digits=3))
```

```
[[32  0  0  2  0]
 [ 5  5  0  0  0]
 [ 0  1  0  0  0]
 [ 4  1  0  1  0]
 [ 5  0  0  1  4]]
              precision    recall  f1-score   support

   DESC:desc      0.696     0.941     0.800        34
ENTITY:other      0.714     0.500     0.588        10
    LOCATION      0.000     0.000     0.000         1
        Name      0.250     0.167     0.200         6
     Numeric      1.000     0.400     0.571        10

    accuracy                          0.689        61
   macro avg      0.532     0.402     0.432        61
weighted avg      0.693     0.689     0.656        61
```

Fig. 8 Logistic Regression classification result.

5.2. Support Vector Machine

Support Vector Machine uses the technique of One-vs-One strategy of binary classification algorithms for multi-class classification. One-vs-One splits a multi-class classification dataset into binary classification problems which splits the dataset into one dataset for each class versus every other class.

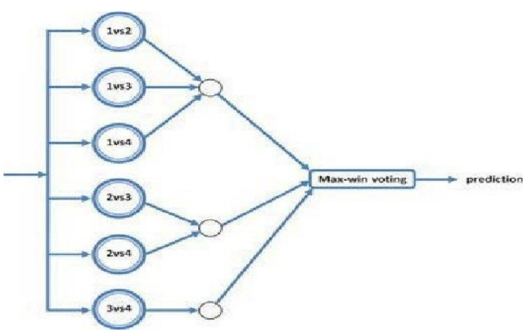


Fig. 9 One Vs One Classifier

It predicts one class label for each binary classification model and the model with the most predictions or votes is predicted by the one-vs-one strategy. This approach is used by support vector machines (SVM).

SVM algorithm has been applied on the supervised data set and for the given new data; it's been classified into the appropriate category. For given 300 questions supervised data set has been applied with SVM and it is giving the accuracy of 70.49%. For K-fold cross validation the test accuracy score mean is 59.04% and it shows the below confusion matrix.

```
#mythreshold=0.5
from sklearn.metrics import confusion_matrix
#predicted= text_clf.predict(X_test)>= mythreshold).astype(int)
cm = confusion_matrix(y_test, predicted)
print(cm)

print(metrics.classification_report(y_test, predicted, digits=3))
[[33  0  0  1  0]
 [ 4  6  0  0  0]
 [ 0  1  0  0  0]
 [ 4  1  0  1  0]
 [ 6  0  0  1  3]]
               precision    recall  f1-score   support

    DESC:desc      0.702     0.971     0.815        34
 ENTITY:other      0.750     0.600     0.667        10
     LOCATION      0.000     0.000     0.000         1
         Name      0.333     0.167     0.222         6
      Numeric      1.000     0.300     0.462        10

     accuracy                          0.705        61
    macro avg      0.557     0.407     0.433        61
 weighted avg      0.711     0.705     0.661        61
```

Fig. 10 SVM classification result.

5.3. K-Nearest Neighbor

Supervised Learning of KNN which is one of the ML algorithms that assumes similarity between new data and available or existing data and classify the new data into the category that is similar to the available category. Here new data is classified into appropriate category by using K-NN algorithm.

It's not learning from training set so it's called as lazy learner algorithm, it performs an action on the dataset. Training phase in KNN algorithm stores the dataset and when it has new data it classifies that given data into an appropriate class/label that is much similar to the new data.

KNN method has been applied on the supervised data set and for the given new data; it's been classified into the appropriate category. For given 300 questions supervised data set has been applied with SVM and it is giving the accuracy of 71.42%. For K-fold cross validation the test accuracy score mean is 67.22% and it shows the below confusion matrix.

```
#mythreshold=0.5
from sklearn.metrics import confusion_matrix
#predicted= text_clf.predict(X_test)>= mythreshold).astype(int)
cm = confusion_matrix(y_test, predicted)
print(cm)

print(metrics.classification_report(y_test, predicted, digits=3))
[[44  3  0  3  1]
 [ 5  9  0  0  0]
 [ 1  0  0  0  0]
 [ 5  1  0  4  1]
 [ 4  2  0  0  8]]
               precision    recall  f1-score   support

    DESC:desc      0.746     0.863     0.800        51
 ENTITY:other      0.600     0.643     0.621        14
     LOCATION      0.000     0.000     0.000         1
         Name      0.571     0.364     0.444        11
      Numeric      0.800     0.571     0.667        14

     accuracy                          0.714        91
    macro avg      0.543     0.488     0.506        91
 weighted avg      0.702     0.714     0.700        91
```

Fig. 11 KNN classification result

5.4 Decision Tree

Decision Tree algorithm is an supervised learning algorithms used for solving classification problems. It creates a training model that predicts the class by learning simple decision rules taking from training phase.

It starts with a root of tree for predicting a class label, it compare the values of the root attribute with the record's attribute and after comparison we follow the branch with the value and jumps to the next node.

Decision tree algorithm has been applied on the supervised data set and for the given new data; it's been classified into the appropriate category. For given 300 questions supervised data set has been applied with DT and it is giving the accuracy of 70.32%. For K-fold cross validation the test accuracy score mean is 61.66% and it shows the below confusion matrix.

```
#mythreshold=0.5
from sklearn.metrics import confusion_matrix
#predicted= text_clf.predict(X_test)>= mythreshold).astype(int)
cm = confusion_matrix(y_test, predicted)
print(cm)

print(metrics.classification_report(y_test, predicted, digits=3))
```

```
[[44  2  0  5  0]
 [ 2 12  0  0  0]
 [ 0  1  0  0  0]
 [ 6  1  0  2  2]
 [ 5  0  0  3  6]]
              precision    recall  f1-score   support

   DESC:desc      0.772     0.863     0.815        51
ENTITY:other      0.750     0.857     0.800        14
    LOCATION      0.000     0.000     0.000         1
        Name      0.200     0.182     0.190        11
     Numeric      0.750     0.429     0.545        14

    accuracy                          0.703        91
   macro avg      0.494     0.466     0.470        91
weighted avg      0.688     0.703     0.687        91
```

Fig.12. Decision tree classification result.

5.5 Naive Bayes

Naive Bayes classifier used for multiclass learning. Trained Classification Naïve Bayes classifiers that store the training data, prior probabilities, parameter values. It uses these classifiers to perform tasks such as estimating resubstitution predictions and predicting labels or posterior probabilities for the new given data.

.

Naïve Bayes algorithm has been applied on the supervised data set and for the given new data; it's been classified into the appropriate category. For given 300 questions supervised data set has been applied with DT and it is giving the accuracy of 65.93%. For K-fold cross validation the test accuracy score mean is 58.22% and it shows the below confusion matrix.

```
#mythreshold=0.5
from sklearn.metrics import confusion_matrix
#predicted= text_clf.predict(X_test)>= mythreshold).astype(int)
cm = confusion_matrix(y_test, predicted)
print(cm)

print(metrics.classification_report(y_test, predicted, digits=3))
```

```
[[50  0  0  1  0]
 [ 7  7  0  0  0]
 [ 0  1  0  0  0]
 [ 9  1  0  1  0]
 [12  0  0  0  2]]
               precision    recall  f1-score   support

    DESC:desc      0.641     0.980     0.775        51
 ENTITY:other      0.778     0.500     0.609        14
     LOCATION      0.000     0.000     0.000         1
         Name      0.500     0.091     0.154        11
      Numeric      1.000     0.143     0.250        14

     accuracy                          0.659        91
    macro avg      0.584     0.343     0.358        91
 weighted avg      0.693     0.659     0.585        91
```

Fig.13 Naïve Bayes classification result.

6. Discussion

We have applied the several classification algorithms for the question data set and the above figure classification result shows there is an increase in the results which indicate SVM, KNN & Decision tree have performed well and showing the good accuracy levels compared to logistic regression, decision tree, and naïve bayes.
After comparing the several algorithms, the accuracy plays a major role to judge the quality of an algorithm to meet the standard value.

$$\text{Accuracy} = \frac{\text{Number of classified questions correctly}}{\text{Total Number of questions}}$$

7. Conclusion

In this paper we have proposed and compared the several classification algorithms to classify the questions by considering the domain specific data set i.e. pomegranate fruit. In future classification task can be done on general data set and on multimedia data sets.

References

[1] https://krishi.icar.gov.in/jspui/bitstream/123456789/4319/2/Bulletin%20English-2.pdf

[2] Kolomiyets, O., & Moens, M.-F. (2011). A survey on question answering technology from an information retrieval perspective. Information Sciences, 181(24), 5412–5434.

[3] Bu, F., Zhu, X., Hao, Y., & Zhu, X. (2010). Function-based question classification for general QA. Proceedings of the 2010 conference on empirical methods in natural language processing. Association for Computational Linguistics 1119–1128.

[4] Bullington, J., Endres, I., & Rahman, M. (2007). Open ended question classification using support vector machines. MAICS 2007 .

[5] Nguyen, T. T., Nguyen, L. M., & Shimazu, A. (2007). Improving the accuracy of question classification with machine learning. Research, innovation and vision for the future, 2007 IEEE international conference on. IEEE234–241.

[6] Li, X., & Roth, D. (2006). Learning question classifiers: The role of semantic information. Natural Language Engineering, 12(03), 229–249.

[7] Alaa mohasseb, mohamed bader-el-den and mihaela cocea "Question categorization and classification using grammar based approach" Information Processing and management https://doi.org/10.1016/j.ipm.2018.05.00.

[8] Muhammad Wasim et al. – "Multi-Label Question Classification for Factoid and List type Questions in Biomedical Question Answering" in 2018 IEEE transactions and content mining volume 7, 2019.

[9] Bastian Haarmann "Mighty Data set for stress testing QA system" in

[10] Hasangi Kahaduwa ; Dilshan Pathirana ; Pathum Liyana Arachchi ; Vishma Dias ; Surangika Ranathunga ; Upali Kohomban

" Question Answering system for the travel domain " 2017 Moratuwa Engineering Research Conference (MERCon).

[11] M.R. Sumalatha ; N.Ahana Priyanka "Analyzing and predicting knowledge of contributors in community question answering services " 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET).

[12] Sharvari Gaikwad; Rohan Asodekar; Sunny Gadia; Vahida Z. Attar "AGRI-QAS question-answering system for agriculture domain" 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI) pp. 1474 - 1478, 2015.

[13] Payal Biswas; Aditi Sharan; Nidhi Malik "A framework for restricted domain Question Answering System" International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT),pp. 613 – 620, 2014.

[14] B. Haarmann, C. Martens, H. Petzka and G. Napolitano, "A Mighty Dataset for Stress-Testing Question Answering Systems," *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, 2018, pp. 278-281, doi: 10.1109/ICSC.2018.00054.

[15] Gautre, T.A. et al. (2018) 'An analysis of question answering system for education empowered by crowdsourcing', Proceedings of the Second International Conference onInventive Systems and Control (ICISC 2018).

[16] M. Devi and M. Dua, "ADANS: An agriculture domain question answering system using ontologies," 2017 International Conference on Computing, Communication and Automation (ICCCA), 2017, pp. 122-127, doi: 10.1109/CCAA.2017.8229784.