AN ENHANCED APPROACH TO IMPROVE THE ENCRYPTION OF BIG DATA BY INTELLIGENT CLASSIFICATION TECHNIQUES

A Thesis
Submitted in partial fulfilment of the requirements for the

award of the degree of

DOCTOR OF PHILOSOPHY (PhD)

in

COMPUTER SCIENCE AND ENGINEERING

By

Gitanjali

41700086

Supervised By

Dr Kamlesh Lakhwani

LOVELY PROFESSIONAL UNIVERSITY
Transforming Education Transforming India

LOVELY PROFESSIONAL UNIVERSITY
PUNJAB
2021

DECLARATION

I hereby declare that thesis entitled "An Enhanced Approach to Improve the Encryption of Big Data by Intelligent Classification Techniques" submitted by me for the Degree of Doctor of Philosophy in Computer Science and Engineering is the result of my original and independent research work carried out under the guidance of my supervisor Dr Kamlesh Lakhwani, Associate Professor, School of Computer Science and Engineering, Lovely Professional University, Jalandhar. This work has not been submitted for the award of any degree or fellowship of any other University or Institution.

Gitanjali
School of Computer Science and Engineering,
Lovely Professional University,
Phagwara, Punjab-144411, India
Date:

CERTIFICATE

This is to certify that the thesis entitled "An Enhanced Approach to Improve the Encryption of Big Data by Intelligent Classification Techniques" submitted by Gitanjali for the award of the degree of Doctor of Philosophy in Computer Science and Engineering, Lovely Professional University, is entirely based on the work carried out by her under my supervision and guidance. The work reported embodies the original work of the candidate and has not been submitted to any other university or institution for the award of any degree or fellowship, according to the best of my knowledge.

Dr Kamlesh Lakhwani
Associate Professor
School of Computer Science and Engineering
Lovely Professional University
Phagwara, Punjab-144411 India
Date:

ABSTRACT

In today's competitive time, data is considered as one of the significant assets and in order to maintain its value, the safety of the data is a major concern. On the other hand, cloud is the technology that has proficiency in storing the data at minimal or no cost at all. In cloud, storing the data in a safe place is one challenge and the data encryption in comparatively less time and storage space is another challenge. These challenges have prevented a number of financial and government organizations to take advantage of its utility. A two- step solution is illustrated in this thesis as a response to the raised issue. The first part of the research work embarks upon classifying sensitive and non-sensitive data. This further ensures that resources are used to encrypt sensitive data effectively. To implement the same, hybrid mechanism is proposed as an algorithm that is based upon Convolution Neural Network with Logistic Regression (CNN-LR). The next phase works on an Encryption Technique that focuses on encrypting the classified data ensuring the time and space complexity. We have used Elliptic-curve Diffie Hellman and Shifted Adaption Homomorphism Encryption (ECDH-SAHE). Lastly, Elliptic-curve DiffieHellman-Shifted Adaption Homomorphism Decryption (ECDH-SAHD) has been used to decrypt the data. The proposed novel approach is entitled Sensitive Encrypted Storage (SES). The results of the proposed model are highly encouraging in terms of efficiency and capability which provides a new dimension to the future researchers.

ACKNOWLEDGEMENT

It is my pleasure to thank all those who have helped me to accomplish this PhD. thesis. Firstly, I wish to express my deepest gratitude to Dr Kamlesh Lakhwani for guiding me throughout this research work.My supervisor has been a continuous source of knowledge, inspiration, motivation and encouragement during the entire course of this research work.

A special thanks to the management of Lovely Professional University for supporting me in the best possible manner and facilitating me in balancing my work and my research. The doctoral programme of LPU has made it possible for me to pursue my dream of research and upgrading my knowledge.

Special thanks to Examiners of end term reports and reviewers of journals who vetted my submissions and gave valuable comments to further improve the work.

I wish to express my profound gratitude to my parents and all other members of my family. Their love, support and unshakable faith in me provide me strength to achieve all my goals in life.

I am also thankful to my husband Himanshu Goel, who has provided me full support during the entire period of my research work. I also take this opportunity to express gratitude to all my teachers who have shaped me and have contributed immensely to my knowledge and skill development since childhood.

Finally I would like to thank each and every person who has directly and indirectly helped and motivated me in this herculean task.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AES	Advanced Encryption Standard
CNN	Convolutional Neural Networks
ECDH	Elliptic-curve Diffie–Hellman
GFS	Google File System
HDFS	Hadoop Distributed File System
IaaS	Infrastructure as a service
KMS	Key management system
KNN	K-Nearest Neighbors
LR	Logistic Regression
MAC	Media Access Control
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
PaaS	Platform as a service
PB	PetaBytes
RNN	Recurrent Neural Networks
RSA	Rivest, Shamir, Adleman Encryption
SAHE	Shifted Adaption Homomorphism Encryption
SAHD	Shifted Adaption Homomorphism Decryption
SVM	Support Vector Machine
SaaS	Software-as-a services
TB	TeraBytes
XML	EXtensible Markup Language

CHAPTER 1

INTRODUCTION

With the rapid growth of internet-based technologies, an immense volume of data has amassed in social media communications, blogging, e-commerce, and surveillance. The biggest challenge here, is to define the unstructured, semi-structured, and structured data separately. Big data analysis addresses huge possibilities of insights. A combination of data mining, text mining, web mining, and NLP i.e. Natural Language Processing techniques id used for Big Data to be analysed in various day-to-day real-life applications. The volume of data is increasing from multiple sources of social media such as Facebook, Twitter, Amazon, YouTube, etc. The enormous amount of sensitive data is associated with these social media platforms and it is quite challenging to analyse all of them. Innovative approaches are needed to separate sensitive data from non-sensitive data. Most of the web facilities like news reports, e-commerce websites, social media networks, blogs, and forums help to express opinions. They can be utilized to extract useful and sensitive information from all the data and further, only useful data can be encrypted or protected, rather than protecting both sensitive and non-sensitive information, which can be a very costly as well as time-consuming process. This research work has mainly focused on the analysis of big data from social networking sites like Twitter to extract useful information with maximum accuracy using various Machine Learning and Deep Learning techniques.

1.1 Big Data: Overview

The gradual directorial change in social networking websites, sensors, feature-packed mobile phones, and cloud computing makes data analytics a complex subject to handle. Abundant procurement of data from multiple resources shunned the petabyte range. The inevitable process of decomposition of these large amounts of data has become a reality in almost all business firms at the current time.

"Big data are the information asset of high volume, velocity as well as variety which need some latest processing forms to allow strong and influential decision making, process optimization as well as insight finding", as mentioned by Gartner. "The collection of large and complicated data that couldn't be processed by ordinary applications of data processing is called Big Data". It accommodates enormous data, which are diverse in nature and spawned at high speeds. As data is procured at a fast pace and in huge volumes, there is a requirement for modern methods and applications to process and regulate the data. In the olden days, certain practices were prevalent in capturing or storing data through various methods like carving on stones, woods, sheets of metals, etc. In subsequent periods, data were netted into a paper cloth and other such storage material of which punching machines became a tool to do so. These developments caused further inventions and the result was floppy discs, hard disc drives, USB drives, etc. Every generation has witnessed a prominent upsurge in quantifying the data stored, and the current generation is having the most potential cloud infrastructure, which can store unlimited data without any hassles. At present, not only the data generation but its capturing and storing is done up to Terabytes (TB) and Petabytes (PB). In the world's scenario of big Data, over a trillion Gigabyte (GB) of new data will be available globally and by the year 2020, approximately 45 Zeta bytes generates almost 7 TB a day.Moreover, 10 TB of data per day is generated by another popular website Facebook. When such procurement of data is there, the task to deal with it, is more challenging than ever. Its rapid growth is shown in Fig. 1.1. Big data is the term that broadly refers to datasets that, which are large and complex in nature , when compared in terms that prevailed inadequate traditional data processing applications.

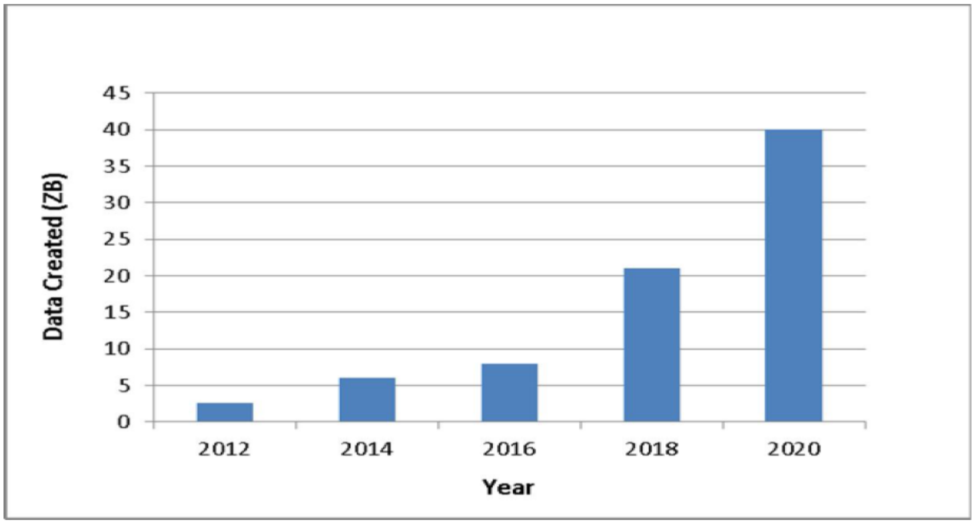


Figure 1.1: Data Growth Rate in Current Decade

At present, the crisis with data analysis, is not only extensive data but the dynamism in data streams and the composition of distinct data types. In fact, the exclusive nature of big data is due to its gigantic size, high dimensionality, diversity, intricacy, unstructured, and imperfect data characteristics that lead to some modification in the statistical and data analysis methods. Though there is a possibility is there to gather more data to discover more valuable information with Big data, it cannot be strongly proclaimed that the information collected is useful in all situations. Sometimes, it carries even more ambiguous and irregular data. For example, a customer may have many accounts which result in lowering the correctness of the analytical outcome.Therefore, many revolutionary data analytics topics are emerging, such as privacy, protection, storage, fault tolerance, and data quality.

There are difficulties related to evaluating, recording, sorting, searching, exchanging, storing, transmitting, visualizing, and privacy information. Too often, the word speaks with the use of predictive analytics or many sophisticated approaches to derive data value, but it never restricts itself to dealing with a specific size of the data collection. It is important to remember that data analysis is about finding fresh patterns, detecting market trends, disease prevention and crime-fighting. Some of the stakeholders, such as scientists, media networks, advertising companies, and governments are confronted with these outsized data sets, particularly in areas such as Internet search, finance, and business informatics.

Big data are considered as the ones that comprise structured, unstructured, and semi-structured data. Leading one amongst these is the unstructured data with an estimated quota of 95% in Big data. Structured data are arranged systematically for retrieval, manipulation, and analysis. What can be seen in relational databases is an example of this. Semi-structured data is not been placed in relational databases, yet its organizational properties allow their efficient analysis. With a few undergone changes, semi-structured data can be acknowledged many times in relational databases. One such example for semi-structured data is XML. Unstructured data never follow any specific format and is largely deprived of meta-data; such as data available from various social media, emails, chats, images, audio and video files, etc.

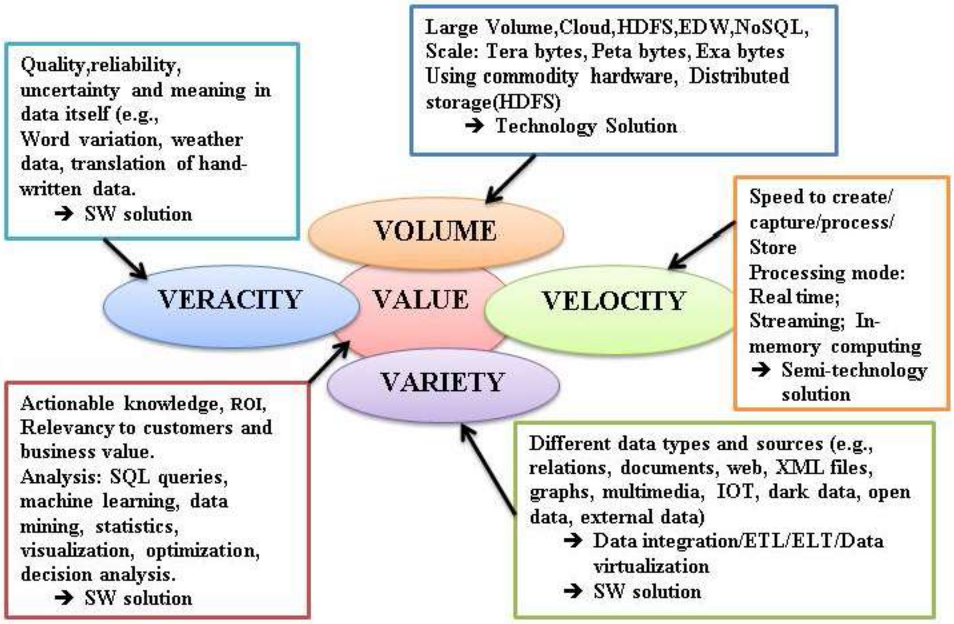

Large Volume,Cloud,HDFS,EDW,NoSQL,
Scale: Tera bytes, Peta bytes, Exa bytes
Using commodity hardware, Distributed storage(HDFS)
➔ Technology Solution
Quality,reliability, uncertainty and meaning in data itself (e.g., Word variation, weather data, translation of hand-written data.
➔ SW solution
VOLUME
VERACITY
VALUE
VELOCITY
VARIETY
Speed to create/ capture/process/ Store
Processing mode: Real time; Streaming; In-memory computing
➔ Semi-technology solution
Actionable knowledge, ROI, Relevancy to customers and business value.
Analysis: SQL queries, machine learning, data mining, statistics, visualization, optimization, decision analysis.
➔ SW solution
Different data types and sources (e.g., relations, documents, web, XML files, graphs, multimedia, IOT, dark data, open data, external data)
➔ Data integration/ETL/ELT/Data virtualization
➔ SW solution

Figure 1.2: Fundamental Characteristics of Big Data

Scientists have encountered limits in a number of areas, including genomics, meteorology, connectomics, complex physics simulations, and biological and environmental science. Big data is the new buzzword in technology.

Its characteristic features can be identified with five 'V's as shown in Fig. 1.2 and Fig. 1.3.

- Volume: This refers to large size of data. The data exceeds terabytes which include records, archived documents, files, transactions, tables, etc.

- Velocity: This indicates the high speed at which the data is procured. It engrosses itself in batch processing, real/near-time data, stream data etc.

- Variety: It denotes various kinds of data (structured, semi structured and unstructured). The data are subjected to multi-factors and influenced to chance variations.

- Veracity: The perseverance of the data with reference to exactness or precision. Some of the key features of the data are genuineness, origin, accessibility, eminence and liability.

- Value: Just the possession of big data, reaps nothing unless it is added with some value. Its importance is shown in Fig. 1.4. The value that is attributed may be statistical or hypothetical which necessitates events and correlations.

The Big data may be developed by a hand-held tool, IoT, multimedia, social network and newly introduced applications. To understand it more clearly, data are adequate to be handled by a single machine and the one that cannot be managed by a single machine is recognized as big data. Whatsoever, the data must be dispensed to operate it. Big data analytics is the procedure through which the analysis of abundant data sets with multiple data types is performed. It is about exploring patterns that are hidden, unknown correlations,client priorities, market trends, and other valuable business information. It could be seen as an extension of data mining.

The Hadoop (Apache Software Foundation) open-source project is the first Big Data analysis platform developed to process computer clusters that are comprehensively distributed, with high performance.

MapReduce, the most widely appreciated function, is fundamentally, a two-stage fault-tolerant analytical routine. Data and the process at hand are initially supplied to different compute nodes and at a later

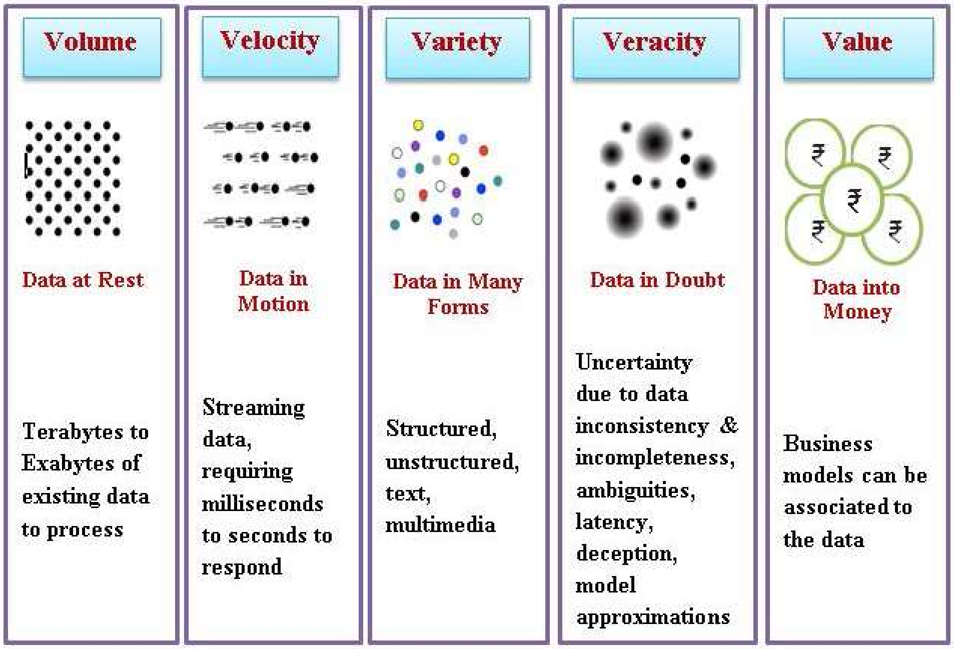

Volume
Data at Rest
Terabytes to Exabytes of existing data to process
Velocity
Data in Motion
Streaming data, requiring milliseconds to seconds to respond
Variety
Data in Many Forms
Structured, unstructured, text, multimedia
Veracity
Data in Doubt
Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations
Value
₹
₹
₹
₹
₹
Data into Money
Business models can be associated to the data

Figure 1.3: Meaningful Representations of 5 Vs of Big Data

stage, the obtained results are consolidated. This is achieved using the Hadoop Distributed File System (HDFS) supported by the (distributed) Google File System (GFS).

IBM's BigInsights and InfoSphere are some of the known commercial platforms for analysis of Big Data and in streams in the order already mentioned. A survey concerning the Big Data's role in interpreting technologies, platforms, applications, and challenges with possible interpretations was furnished about designing Big Data systems.

1.1.1 Challenges of Big Data

Big Data (BD) has a vast part to play with new technology trends in this world. Big Data is contributing to unstructured data, which is very difficult to process and analyse. The speedy growth in unstructured data faces several challenges that are implied in academics, industries, and organizations. Primarily, BD challenges include analysing data, sharing of information, storage of information, query updating information by changing the way to transform.

1.1.2 Big Data Analytics

"Big Data Analytics (BDA) is the method of analyzing large data sets that contain a variety of data types, i.e. big data to identify hidden patterns, unknown associations, market dynamics, consumer desires, and other useful business details." Big data analytics are useful for companies to promote, scale and expand their business.

It helps the companies to understand the pattern of sales, profits and even in internal functioning of their organizations which certainly takes the companies a few step closer to achieving their desired targets.

The following five types of big data analytics:

- Prescriptive Analytics- It is one of the most popular and effective ones. Under this, the analyst comes up with a number of recommendations for any particular problem or a situation currently in hand out of multiple options which are available. The choices are made in such a way that the decision helps in mitigating any risks in the future or can ease any advantage resulted from the decision.

- Diagnosis Analytics - As the name suggests, this technique is helpful when a problem has to be diagnosed to know the reason behind it. Therefore, the approach of this technique is backward in nature. It helps immensely while understanding the pattern or behaviour of the old customer base.

- Descriptive Analytics - This one is similar to diagnosis in the way that it also looks backwards at the data. The difference between the two is that the Descriptive Analytics works on What Happened whereas Diagnosis works on Why it happened. Segmentation and Pattern Identification are majorly covered under this category.

- Predictive Analytics It has a futuristic approach as the name tells. This includes the processes which tell how much is the possibility of an event to occur in the future.
- Outcome Analytics- Outcome analytics completely revolve around the customers, their purchase behaviour and the likelihood of opting for your products in the future.

The following four types of big data analytics need to be considered:

1. Descriptive details (what happened)
2. Diagnosis (why did it happen)
3. Predictive (what's going to happen))
4. Outlook (what action to take)

1.1.3 Big Data Analysis

Big data analysis has an important part to play in the big data value chain. It may give useful values through decisions, recommendations, supports, or picks. Therefore, data analysis encompasses a wide range of packages. Research can be divided by Big Data Analysis into five key technical fields:-

- Structured analysis of data

- Review of text data
- Review of Internet data
- Multimedia analysis of data
- Smartphone review of data

Analysis is the process of breaking a complicated subject into smaller parts to get benefit from it, analytics is the technology of analysis. Basically, analysis and analytics perform an identical feature, but in experience, analytics is the function of technological know-how. According to International Data Corporation (IDC) widely used big data analysis techniques are as follows:

- Learning the rule of association
- Analysis of Tree Classification
- Algorithms to Genetics
- Using the Machine Learning
- Analyzing Regression
- Analyzing Sentiments
- Social network analysis

From the different fields of analysis, this project work only focuses on the sentiment analysis of big data. The prerequisite knowledge to work with sentiment analysis on social media texts needs the following research fields:

- Text Data Analysis: For natural language processing

- Web Data Analysis: For extracting the web content.

This work primarily focuses on classifying big data with maximum accuracy into sensitive and non-sensitive social media data.

1.2 Social Media Text Analysis

Social media has greatly revolutionized this digital era. The technological improvement has brought people closer virtually. The exchange of knowledge and communication between people is easier. There is an ample number of social media networking sites available in the market today that come with different flavours and features. The giants in social media would include Twitter, Facebook, Instagram, Google Plus, and Pinterest. Each one is different in its own way. Among them, Twitter is usually the primary platform people use to voice their opinion on any topic. There is a huge growth and demand for the online information derived from Twitter.

The amount of reviews has been increasing day by day at an enormous speed, making it more subjective and opinionated. When it comes to posting on social media platforms, each user tries to cover the reviews on any situation or product in a very detailed manner. These posts help an organization derive valuable business insights. Analysis of these posts is a more productive way of collecting information that is vital for making decisions. There are innumerable posts on any given topic, growing each minute. Analyzing these data manually and making a decision is a cumbersome process. This calls for a need for algorithmic methods to analyse and display the decision or insight, within this large amount of data. This voluminous and variety of data not only contains valuable insights, but unwanted details like URLs, sarcasm, symbols, abbreviation, other language alphabets, and so much more, leading to noisy data. Sentiment analysis is the process of extracting information from this noisy data. Hence, sentiment analysis techniques are used to simplify the process of classification.

Analysis of sentiment aims at defining the writer's attitude concerning some subject or meaning. This technology's most important task is to discover the emotions hidden in the subject. Sentiment Analysis is the thorough study of recognizing the user's thoughts, emotions, state of mind, and point of view on an occasion in natural language. Recent events prove that this work has come to an overwhelming level. For distinctive classes and points, the study may outperform the classification of positive versus negative polarity, and handle the entire field of actions and feelings. In the field of sentiment analysis for predicting social attitudes using diverse structures, a significant measure of research has been conducted.

An essential and typical task in sentiment analysis is polarity classification. Primarily, there are positive and negative classifications, but there is also a third classification known as neutral polarity. The sentiment of individual words has been utilized by most sentiment analysis strategies. Lexicons are word and expression arrangements named either with their earlier polarity (neutral, negative, or positive) or numerical ratings.

The approach which is being used as dictionary-based approach begins with a seed-word structure. These seed words are collected from various sources and stored in a dictionary. These words are then expanded to the remaining words and expressions in that dictionary (for example, WordNet) in the light of the synonym and antonym structures of the dictionary. Contrasted with the manual analysis, this dictionary-based approach often has a higher scope yet brings down precision. The corpus-based approach is another methodology that focuses on creating a domain-dependent lexicon using a domain corpus.

1.2.1 Applications and Tools for Text Analysis

Social media text analysis is also called sentiment analysis. Wide range of applications for analysis of sentiment from social media text are available, few applications are mentioned below.

- Product and Service Surveys - The most well-known use of sentiment analysis is in the buyer products and administration survey field. Numerous sites are giving the products a computerized outline audit. A prominent case of that is Google Product Look.

- Reputation Monitoring - Many social media sites, such as Twitter and Facebook, are a source of convergence for many applications for sentiment analysis. The most recognized application is Twitter to check the popularity of a particular brand.

- Result Expectation - By dissecting sentiments from applicable sources, one can anticipate the plausible result of a specific occasion. For example, an analysis of sentiments can give generous appreciation to hopefuls running for different positions in elections. It empowers campaigning administrators to track voters 'opinions on different issues and to identify the hopefuls' talks and activities.

- Decision Making - Another vital application is that sentiment analysis can be utilized as a vital variable helping the basic lead-

> ership frameworks. For instance, in the money related markets venture, there are different news stuff, articles, online newspapers, and tweets about every open organization. A framework for sentiment analysis can use these different sources to discover articles that further examine the organizations and add to the feeling about them as a solitary score that can be used by a computerized framework for exchange.

Numerous studies have given techniques and apparatuses utilized for opinions examination. The most utilized instruments for recognizing the sentiments extremity (positive and negative effect) of a message is because of the emoticons. These emoticons symbolize dismal or face-based sentiments, even though there are wide scopes of non-facial varieties. Hence, emoticons have been frequently utilized as a part of the other procedures for building the dataset in directed machine learning procedures.

SentiWordNet is a lexical asset openly accessible for supporting assumption order and assessment mining applications.WordNet is also known as an English lexical that gathers modifiers, things, verbs, and so on into equivalent word sets called synsets.

There are no best algorithms or tools to accurately classify the text. Some are best for a particular domain. For example, movie reviews, product reviews, Twitter posts, and topic-specific text. Researchers are trying to find a generic tool or algorithm that accurately classifies the given text. There are a lot of approaches mainly based on machine learning algorithms or dictionary-based ones. Researchers have explored Decision Tree, SVM, Naïve Bayes, or Entropy approach and even some hybrid models have been proposed that combine two or more approaches. On surveying and analyzing various algorithms and working of various models, it is very difficult to identify the most appropriate model for text classification [115, 134].

1.3 Overview of Classification Techniques

There are various classification techniques which are listed below:-
Supervised Classification: This learning is prepared using labelled data, for example, information where the ideal yield is known. Administered learning gives a dataset comprising of the two highlights and marks. The process of regulated learning is to build an estimator that will almost certainly foresee the name of an article given the arrangement of highlights. The controlled calculation gets a lot of highlights as contributions alongside the associated right yields, and the errors are discovered from calculations by comparing its genuine performance and correct yields. It then adjusts the model accordingly [80]

Naive Bayes (Generative Learning): This grouping system is dependent on Bayes Theorem. Basically, it agrees that the particular present in a class is not synonymous with any other item present in any other class. Bayes display is anything but hard to make and especially helpful for extremely large sets of information [9, 39, 105].

$$P(c|y) = \frac{\mathrm{P(y|\ c)P(c)}}{\mathrm{P(y)}}$$

Logistic Regression (Predictive Learning Model): It is a statistical method of analyzing a set of data in which there are more independent variables that de-finish a result. The result is measured with a split variable.

$$\text{Yi} = \beta 0 + \beta 1 \text{ Xi} + \epsilon \text{i}$$

SVM : It is a linear classifier of the non-probabilistic binary. In multidimensional space, an SVM model is essentially a representation of different classes in a hyperplane. In order to minimize the error, the hyperplane will be generated in an iterative way by SVM. SVM attempts to split the datasets into groups in order to find a maximum marginal hyperplane. [128, 151]

Decision Trees: Decision tree is used to construct classification or regression models in tree structure form. A data set is divided into smaller and smaller subsets, whereas a decision tree is incrementally generated at the same time. The final result obtained is a tree with nodes and leaf nodes decided. There are two divisions for the decision node and a leaf node is used to represent the classification.The root node is the decision node that is present in a tree at the top, corresponding to the best predictor. Decision trees can handle both numerical and categorical details. [140]

KNN: In order to store the current cases, this algorithm is based on similarities and utilizes these cases for future classification. It is used mainly in statistical estimation and identification of patterns. It classifies the data according to the class of the nearest neighbour. [13]

Random Forest: Random forests or random decision forests are methods for arrangement, recurrence, and various assignments that work by building a multi-study of choice trees to prepare time and yield the class that is the grouping or recurrence of the trees. The irregular choice is right for the propensity of decision trees to over-fit set preparation [123].

$$\mathrm{H} = -\mathrm{p(x)logp(x)}$$

Neural Network: : A neural network consists of neurons that translate a data vector into a certain yield. The data is taken up by each neuron, which handles it with ability and then passes the yield on to the next layer. The structures are described as feed-forward: on the layer below, a neuron reinforces its contribution to each of the neurons, yet there is no feedback to the previous layer. Weights are connected to the signs passing starting with one neuron then onto the next. The grouping is a directed learning approach that is utilized for AI and measurements. The characterization is finished by using the learning idea that depends on the information input given to it. Information is grouped by it based on bi-class and multiclass, like sex order in male or female characterizes the messages in the spam or not spam. These approaches are mainly used to classify the documents, biometric identification, speech recognition, and hand-sentiment analysis by learning non-linear models [82].

Semi-Supervised learning is a system that utilizes unlabeled information for preparing. Numerous analysts have discovered that unlabelled information, when utilized related to a little measure of named information, can create considerable improvement in learning exactness over unsupervised adapting, yet without the time and costs required for administered learning. The expense related to the naming procedure may render a completely named preparing set infeasible, though securing unlabeled information is relatively reasonable. In addition, semi-administered learning is a hypothetical interest for Artificial Intelligence and a model for human learning.

Unsupervised Learning uses information that doesn't have authentic names and the goal is to investigate the information and discover similarities between the objects. It is the system by which the information itself finds marks. For example, unsupervised learning works admirably on transactional knowledge, identifying fragments of clients with comparable characteristics that could then be dealt with similarly in supporting efforts. During such learning algorithms the detail brings in a few highlights. Recently learned highlights are used to interpret the class of information at whatever point new knowledge is provided. It is used primarily for grouping and emphasizing reductions. Deep Learning and neural networks are also utilized for the processing of the natural languages which help the machine to understand the natural languages used by human beings to take commands or queries and perform the multiple tasks given to the system. In this process, multitasking learning is used as the input undergoes six different tasks in which its syntactic role is checked. Each word is given a unique tag, atomic elements are labeled, and the language model is checked with semantics related words. After analyzing huge databases, one general neural network architecture for natural language processing is made in which simultaneous learning task is done to improve the performance while various tasks are applied. Deep learning and neural networks are utilized in various other things like pattern matching and machine learning.

RNN(Recurrent Neural Networks) RNN is a network of neuron-like nodes, each with a one-way connection to every other node.In RNN, ht, which acts as a network memory and learns contextual data that is essential for natural language classification, denotes the hidden state.

At each step, the output is determined on the basis of the ht memory at time t and the current input text.The key feature of an RNN is its hidden state, which captures the information's sequential dependency. A specific type of RNN capable of recalling information over a long period of time is the Long Short Term Memory (LSTM) network.

CNN (Convolutional neural networks) CNNs are a form of neural networks that include layers called convolution layers that are capable of interpreting specific data. A convolution layer has a number of filters or kernels which it learns from the data to extract different feature types. The kernel is a 2D window, sided over the input data that perform the process of the convolution. In our experiments, temporal convolution has been used which is appropriate for the study of sequential data such as tweets [134].

1.3.1 Comparison between Various classification Techniques

This section comprises quantitative and qualitative analysis of 61 papers. The fundamental contribution of this research work relies on data-based classification using best machine learning classification approach. The classified technique with maximum accuracy and precision will be analysed.The analysis of existing classification techniques has been demonstrated in the below Table 1.1. The aim is to design an intelligent classification techniques for reducing the error in the prediction of the sensitive data. A few calculations or mix of calculations as hybrid methods have been proposed for programmed classification of sensitive information, among these calculations. NB, KNN, and SVM classifiers have appeared suitable in the current writing.

Bolster vector machine calculation has been recognized as a stand-out amongst the best content characterization strategy in the examinations of directed machine to speed up the calculations. SVM with its Structural Risk Minimization standard limits the upper bound on the generalization mistake. Be that as it may, SVM has discovered a few troubles in parameter tuning and bit determination. On the off chance that a reasonable pre-handling is utilized with k-NN, at that point this calculation keeps on accomplishing extremely great outcomes and scales up well with the number of records, which isn't the situation for SVM. NB accomplished great execution with pre-handling. k-NN calculation performed well as increasingly neighbourhood normal for archives are measured, anyway the arrangement time is hard and long to discover ideal estimation of k.

Table 1.1: Comparison of various existing classification techniques

Ref. No	Algorithm	Findings	Limitations
[9]	C4.5 Algorithm	Simple to execute and easy interpretation of models.	Doesn't function admirably on a huge set of trained data and little variation in information can prompt diverse decision trees.
[39]	ID3 algorithm	More accurate and increased detection rate and less space consumption.	Requires huge measure of memory to store tree and requires large searching time.
[123, 128]	Rocchio's Algorithm	This type of algorithm is simple to execute, effective in computation	This algorithm is low in classification accuracy.
[113]	Decision Tree	It can easily deal with Noisy or incomplete data.	The classification error rate is high, while the training set is limited in comparison with the number of classes.
[8]	Neural Network	This algorithm has the ability to work with incomplete knowledge and having high fault tolerance.	Major Limitation of neural network is there is no specific rule for determining the structure of artificial neural networks. It is based on experience and trial and error.
[151]	Naïve Bayes	The main limitation of this classification method is its relatively low classification presentation as compared to other type of discriminative algorithms.	It is easy to implement and requires small trained datasets for estimating the necessary parameters for the purpose of classification.
[45, 126]	Multi layer	These are flexible and can be utilized for learning the procedure of mapping from inputs to outputs.	Multi layer preceptor cannot guarantee that the minima it stops at during training is the global minima.
[84]	CNN	The advantage of utilizing CNNs is their ability to advance an internal representation of a 2D image.	For non-image data, CNN is not good as CNNs achieve advanced consequences on complications such as data classification utilized in SA and associated difficulties.
[85]	RNN	RNNs recognized the success when employed with arrangements of paragraphs and words generally called NLP.	Recurrent neural networks are not good for tabular datasets and for image data input

1.4 Overview of Encryption Techniques

Cloud computing reflects the way data security is provided, the way the data is securely transferred from user to user, the way the data is securely stored, and the way the secured data is processed within the infrastructure itself. The cloud computing architecture includes a stack of three layers: resources, platform, and applications. Out of which major concentration is on resources wherein how data is secured and securely accessed from the cloud. Cloud is a resource area managed and maintained by a third party via the internet. Usually, in the past, the enterprises had their own data storage areas in their own locale due to which a heavy maintenance cost used to be a burden to the company. Later, the advantage of the third-party space provider was taken by many of the enterprises to store their data. This third-party space provider is the cloud which is composed of hardware, network, storage space, interfaces, and mainly the services using which any user can access the cloud. With the advent of the third-party service provider, the question of secured data has raised which has deviated the cloud to another big area of cloud data security. This is the major concerning factor as a part of cloud storage services. Not only enterprises, today, but many data users are also running behind cloud services; as it provides optimal and unlimited data storage for the user to access his private data from multiple devices (e.g. Tablet mobiles, smartphones, laptops, etc.). With huge weightage on the cloud, it is expected by enterprises and users that the cloud providers may provide more reliable data security services like data confidentiality, data authentication so that their data can be secured within the cloud. With this, cloud providers are increasing the complexity of their services. These services from the cloud have completely changed the way enterprises manage and access the data outsourced at the cloud. In the past, the most reflecting factor from the user and the enterprise is whether data stored with these services of service providers results in confidentiality, or whether other un-authenticated users have the access to the confidential data. Hence, availing of cloud data storage services for storing the most valuable and confidential data depends primarily on whether the service provider can offer high security and assurance to meet the desired requirements. Noticing the way how most cloud data storage infrastructures are designed and constructed, we can find that these data storage services do not provide enterprises and users with sufficient levels of security which leads to a high risk of user's private data from external and internal attacks. These attacks are like: data exposure (no data confidentiality) and data accessing (lack of proper authentication mechanism). To address the above security issues, confidentiality and data access controllability with a strong cryptographic guarantee should be maintained [68, 146]. The thesis presents a strong data encryption mechanism for data confidentiality. The thesis makes use of a novel approach of classifying the user data to fall into various classes like confidential, less confidential and public to raise the strength of security parameters over these class data. Today, most advanced computing technology has made many enterprises outsource their private data and computations to cut off the burden of many economical storage problems. The technology has paved way for two cloud infrastructures: the private cloud and the public cloud. The real data owners maintain and manage the private clouds wherein they restrict the data to be accessed by a few authenticated users of their interest only. The public cloud, on the other hand, is managed and maintained by a third-party cloud service provider and here the owner's data is completely out of his control and potentially can be used by many unknown users. With public clouds, trust (which is feebly called security) is the most concerning factor that can be alternatively achieved by defining the best security specifications. With extending business expansions, public clouds are preferred. Thereby, adequate concentration is on enhancing the security of public clouds. Two major and essential security services to be provided to maintain data security in the cloud are explored in this section i.e. Encryption time and Storage Space. Here, strong data encryption mechanisms are needed. For efficient data accessing many of the cloud service data partitioning mechanisms have been implemented by providers. Partitioning is essential as in its absence the whole of the owner's data is encrypted and stored in the cloud. As the whole data is encrypted, while decrypting at the receiver end the approach failed to raise the data owners' trust as the owners feel that decryption may lose their sensitive data which is a part of the whole data. As a promissory move of increasing this trust, what followed by the cloud service providers is data classification into partitions i.e. allowing the data owner to classify his data to fall into various classes like sensitive and public. Only the owner can access confidential data and all authenticated users of the cloud can access public data. With such data classification, the most concentrated factors from the cloud are:

1. What is the strength of data confidentiality and how users are authenticated to various classes?

2. What is the strength of the keys that are generated to access various class data?

1.4.1 Key Management

The most critical phase of any security system dealing with data encryption is key management[21, 41]. The necessary cryptographic safety specifications and key lifecycle management policies must be defined. Usage and access to these cryptographic keys should only be given to the approved users and those keys should be revoked when users no longer need access to resources. Cryptographic keys used to encrypt the data are to be produced in a secure and complex manner.The key generation process may involve the use of a strong Random Number Generator (RNG). The keys should never be transmitted in their original form after generation and should be wrapped inside a secure element such as a smart bix or some other hardware service. Initially, various ways of managing and storing encryption keys should be prepared. For instance, encryption keys are to be stored on the same data server or another server or handled by a third-party provider of such services are three different ways. Managing and hosting a Key Management Service (KMS) can be too expensive for a cloud provider in a multi-tenant cloud environment. As a result, cloud providers' keys can be protected with software-based solutions that do not meet the requirements of physical security. Network Security has a lot of policies that stop hackers from breaching the information system and causing damage. Data and message protection is the key concern of network security as it is being transmitted over the computer network to the recipient. Network protection ensures that data and information arrive at the destination promptly and in good order. IT practitioners are now getting more concerned with data and post-safety. The Cryptography software was developed to allow users to freely communicate on a computer network.

It is a network where the transmitter sends the receiver an authenticated message. The receiver has to restore the message to plain text. That would be a plain text message to be sent. A code known as ciphertext is encrypted by the transmitter and sent to the receiver over the network. The third-party is attempting to manually push the machine to preach a message that would be of no interest to him. Encryption keys are a must for the receivers. Cryptography is an important piece of equipment used in cyber security. This provides the four most critical information security components in organizations

1) Privacy – Encryption system will protect information and communication from unauthorized disclosure and access to information.

2) Data Integrity – The cryptographic hash limits are supposed to play a fundamental role in ensuring that customers are protected by data.

3) Non-repudiation – The modernized mark gives the non-disavowal party the opportunity to plan for the verbal conflict which may grow as a result of the sender's dissatisfaction with the passing message. Each of these important cryptographic organizations has engaged the business in an extraordinarily lucrative and intensive way that uses PC systems to lead over the frameworks. Nowadays the devices have gone all over the world and information has taken on the bits and bytes of the mechanized kind.

Exhibit Day Cryptography provides an efficient technique of action to ensure that the adversary's destructive points are smashed while simultaneously ensuring that true blue customers get to the data. Encryption is an oversight mechanism that defends against passive behaviour. Security is known as messages of authentication which is often called

Symmetric Encryption, Standard Encryption, or Single Key Encryption.

Cloud computing is a technology that delivers services over the Internet. For cloud computing, people use so many tools using remote services over the network. This means that the consumer ends up having minimum requirements but using the intense computing capacity. Cloud computing provides a centralized location where all of the hardware and software resides if anyone wants to use it. All the data is then placed at some different places on remote servers. Therefore, this data must be protected from unauthorized access, alteration, and attacks. Cloud Data Security means protecting information from attacks in the calculations. Three core goals of cloud data security are :1) Availability 2) Confidentiality 3) Integrity.

1.4.2 Security Issues

Cloud storage has various security problems. These problems are discussed in this section.

- **Data Issues:** Cloud data can be accessed from anywhere and everywhere, this data can be general, private, and confidential. Data stealing is one of the biggest problems because data can be accessed and modified by cloud service providers or consumers. Data loss will occur if service providers shut down their services due to some legal or financial problems.

- **Compliance:** It is the association's responsibility to work in compliance with relevant regulations and requirements. Many companies face compliance issues which are the location of the records.

- **Malicious Insiders:** They are company employees. These are appointed by cloud service providers to operate and administer cloud services. But, the organization's sensitive data in the cloud can be stolen, or can also be shared with other organizations.

1.4.3 Types of Encryption

Encryption is the process of converting plain text into cipher text which ensures data privacy. Various types of encryption exist:

Symmetric Encryption: This is also called 'Encryption of the Private Key'. The same key is used for encoding and decoding in Symmetric key encryption. The key shall be distributed before the data is transmitted. A brief description of some well known symmetric encryption techniques is mentioned below. Also the comparative analysis of existing Symmetric Encryption approaches i.e. Blowfish, DES, 3DES, AES has been demonstrated in Table 1.2

- **3DES** It stands for standard triple data encryption. Essentially it's the DES modification. It is DES replacement due to advances in key searching. 56-bit key is not enough to encrypt sensitive data. 3DES extend the key size three times with three different keys, three times.
- **AES** This reflects Advanced Interest in Encryption. Any mix of data (128-bit) and key lengths of 128, 192, and 256 bits can be recognized. It goes through ten rounds for 128-bit keys, twelve rounds for 192-bit keys, and fourteen rounds for 256-bit keys to provide final cipher-text or restore the original plain-text.
- **BLOWFISH** It is an asymmetrical block cipher and the block size is 64 bit which is used for protection and encryption purposes. It divides up the block into two halves. It is available to all users at no charge. It is efficient for microprocessors, because of its simplicity.

Table 1.2: Comparison between Symmetric Techniques

Parameters	**DES**	**3DES**	**AES**	**BLOWFISH**
Key Used	Same key is used	Same key is used	Same key is used	Key Same key is used
Throughput	Less	Less	Less	Much high
Encryption Ratio	High	Moderate	High	High
Tunability	No	No	No	Yes
Power Consumption	Higher	Higher	Higher	Very Low
Key Length	Key length in DES is 56 bits	Key length in 3DES is 112-168 bits	Key length in AES is 128,192,or 256 bits	Key length in BLOWFISH is 32 bits to 448 bits
Speed	Fast	Fast	Fast	Fast
Security against Attacks	Brute force attack	Brute force,	Chosen plain	Dictionary attacks

Asymmetric Encryption:It is also called' Encryption of Public Key.' In asymmetric encryption, two keys are simultaneously generated. These two keys vary from each other for encryption and decryption. The decryption key cannot be extracted from the encryption key. The message can be encoded (encrypted) by anyone but it can only be decoded by the one who knows the corresponding private key. Example: Diffie-Hellman keys, SSH, SSL, DH, and RSA

- **RSA Algorithm**

It is also known as asymmetric cryptography, which is the public key cryptography. This approach to cryptography is distinct from symmetric cryptography, where only one key is used for encryption and decryption. Two keys are used in public-key cryptography, one to be coded and one to be decrypted. Those 2 keys are known as the public and private keys. The private key is kept safe by the message sender, and the public key is given to all recipients of the message.

- **Diffie Hellman Algorithm**

Key exchange is the basis of Diffe Hellman cryptography. The two parties have to exchange key secrets to encrypt messages. It is based on discrete logarithms ' computational complexity. Diffie-Hellman is based on both the encryption and decryption of symmetric key exchanges.

Hashing: Using this approach that message is encrypted using the mathematical symbols; this is further useful for digital fingerprint. The integrity of the message is maintained and is done on a priority basis. Basically, there are different phases that are involved in the cycle of encryption algorithm. Plain text or the message is used as input in the process of encryption and further this plain text is encrypted with the public key. As an output the cipher text is produced and is further decrypted by the key that is with the receiver and in this way original form of message is accessed. The entire process depends upon the different key sizes that are allocated. The overall process can be understood in this way: Ci= Ek (Pi), Pi = Dk (Ci) The symbols represented above take following form: Pi - plaintext, Ci -cipher text, Ek- encryption method, Dk- decryption method, K- Key

- **Salt Hash Encryption Technique** The World Wide Web has changed a person's life. In one day, more than half of the world uses the Internet and communicates to exchange data and information with each other. But there is a chance of leakage of personal information when uploading such data over the Internet. The user's password can be stolen by cyber stalkers and used to threaten the personal data of the victims. Salt hash technique store passwords and sensitive data using the salt hash method. There are different data protection approaches, but in some applications, salt can prove dominant, among others.

1.5 Thesis Organization

This thesis is devoted to design an enhanced approach to improve encryption of big data by intelligent classification techniques. The chapter-wise organization of further report is mentioned below.

Chapter 2: Related Work

Extensive literature survey related to the bigdata text classification, encryption, and sentiment analysis of social media text has been done. Moreover, research-gap, problem statements and research objectives have been finalized.

Chapter 3: Sentiment Analysis and Homomorphism Function Properties

To implement and apply the proposed research, sentiment and opinion analysis of social networking data has been chosen. In this chapter homomorphism function properties of data used for sentiment analysis has been discussed extensively.

Chapter 4: Sensitive Encrypted Storage

In this Chapter, an optimal solution and methodology to classify the social networking text into sensitive and non-sensitive segments is proposed. Moreover in this chapter the accuracy of various classification algorithms is compared and the best one is chosen for the further process. Further, only the sensitive data is encrypted so as to save time , space and storage. The detailed contribution is discussed in Chapter 5.

Chapter 5: Result and Discussion

Proposed classification and encryption technique has been executed on described dataset. The classification accuracy, encryption and decryption time of proposed technique is compared with well known existing techniques is included in this chapter.

Chapter 6: Conclusion and Future Work

The research work hereby concluded in this chapter, emphasizing the contributions made towards the proposed research domain and presenting future directions in this research area.

CHAPTER 2

RELATED WORK

Outline

This chapter incorporates an extensive literature survey related to the Big Data text classification, encryption, and sentiment analysis of social media text. Moreover, research-gap, problem statements and research objectives have been finalized.

2.1 Review of Classification Techniques

An approach of data classification based on data confidentiality was presented by Zardari et al. [159].To identify the data according to security needs, the KNN algorithm is used. It classifies the data in a non-sensitive and sensitive type that presents the data with a need for protection. For encryption, the RSA algorithm is used and is simulated with the CloudSim Simulator[51].

A variable data classification index was proposed by Moghaddam et al. [102] to provide privacy and security to the cloud data. The value of the index is determined by using different parameters and the main parameters are confidentiality, integrity, and availability.

A deeper discussion was done by Zardari et al. [159] on data confidentiality and data retrieval problem in cloud computing. These issues were solved by using the classification of data and the cloud model. The problem was solved by using the hybrid multi-cloud model with data classification. This model was based on multiple clouds, classification, and different numbers of clusters.

A model based on the classification and secure cloud computing was presented by Tawalbeh et al. [32]. The overhead and processing time used in the security mechanism is minimized by this model. It defines the security at a different level with variable key sizes. The proposed model is tested with different security mechanisms and it gives effective outcomes with high efficiency in the proposed work.

A classification method was proposed by Shaikh et al. [133] which works on different parameters. These parameters define the different dimensions. Depending on the level and the protection needed, data security may be offered. Data leakage and privacy security are covered by the proposed system.

The K-nearest neighbour classifier for providing data confidentiality in the cloud-based data was proposed by Zardari et al. [159]. The technique is extended to the virtual cloud and the information is categorized according to its security needs. KNN classifier classifies the data into two classes that are sensitive and non-sensitive data.

The assessment of the customer's opinions by the service of microblogging for example Twitter was portrayed by Marvin et al. [53]. Twitter is internet-based app on which every single individual communicates their perspectives. The opinions of the customer are the basis of examining the consumer perception about the individual item. On this purchaser, view performs aspect-based SA examination by PoS labelling and, parsing dependency from Natural language handling. It extracts the positive, negative, and neutral aspects from tweets. In the proposed method, the programming toolbox is de-marked so that it first performs the tweet extraction, then it filters subsequently while it later analyses the polarity of sentiments and after that shows the outcome.

The current work of sentiment mining based on the word level, not on the sentence level was presented by Aung et al. [10]. This finds the unequivocally communicated opinions. The paper proposed an approach that operates on the trained set of data that analyses and provides positive, negative, and neutral surveys for various items. ABSA i.e. Aspect Based Sentiment analysis operates on the distinctive aspects of the substance and which in return demonstrates the polarity. To execute ABSA; NL and ML are utilized. The informational index utilized in this proposed paper has 845 aspect-based category annotations in the test data and 1654 aspect category-based annotations in a trained dataset. The performance of programming is estimated by the logistics regression algorithm and SVM.

A novel technique as proposed by KeumheeKang et al. [75] for distinguishing the clients with burdensome states of mind by examining their tweets on daily basis for a significant stretch of time. All media sorts of tweets are exploited, i.e. pictures and emojis along with texts. To survey the legitimacy of the proposed technique, two sorts of examinations were carried:

1) The proposed multimodal investigation was tried with various tweets, and its exhibition was contrasted with SentiStrength;

2) It was applied to arrange 45 clients' psychological states as burdensome and non-burdensome ones.

The test results affirmed that the proposed multimodal investigation technique has a higher exactness than existing strategies and it can foresee people's states of mind all the more proficiently.

A method was discussed by Rongrong et al. [22] which was based on the visual SA methodology. In this, an overview is introduced which characterizes various methods utilized for the visual SA investigation. In this kind of investigation, pictures are utilized to foresee the feelings of the person. The study fundamentally centered around the front-line techniques that are utilized in the picture examination process. This review depicts the new stage for specialists because initially the research is done on the content yet visual opinion ontology is another idea for doing something else. For successful visual SA investigation, the idea of profound learning is helpful.

A supervised learning algorithm was presented by Turney et al [149] which characterizes the survey that is approval and disapproval (thumbs up and down). The normal semantic orientation is utilized to anticipate the classification-based survey. The association based on the positive and negative relationships with the audit demonstrates the direction of the survey. The semantic direction is determined by utilizing PMI-IR which is the center advance of this examination. The proposed calculation gives distinctive exactness on various kinds of tweets like on motion pictures 74%, banks and autos 80%, and 84% on travelling audits.

Enormous work was done by R. Socher, et al. [136] on the sentence level expectation of label-based technique. This utilized another methodology that depends on the prediction based on sentence-level in re-cursive kind of autoencoders. The suggested work is conducted on the standards of sentiment change and sentiment lexica. The dataset used in this study is focused on real-life client stories that included feedback on different labels and was collected using multinomial conveyance to capture enthusiastic responses.

A method based on the unsupervised data extraction technique was presented by Ahmed et al. [5] which is utilized to remove the sentiment or opinion from the audits. This work is carried out in the order depicted in the steps below. Initially, the product characteristics are recognized and, secondly, the object-based assessment from which the item is differentiated is defined and, thirdly, the extremity of the feelings is evaluated. The last advance of the proposed strategy is positioned dependent on their quality. The semantic direction is controlled by utilizing the relaxation labelling methodology. The results are based on the precision and it recalls the review after effects of the proposed approach which demonstrates the efficiency in sentiment identification.

A focused work at the standard Arabic information for sentiment analysis was done by M. Abdul-Majeed et al. [1]. A dataset is gathered in this work and afterwards, the automatic classification step is begun in which tokenization is done on the information. The process based on a two-stage grouping procedure is performed on the dataset. The after effects of the proposed methodology demonstrates the adequacy of the methodology.

A huge work at the visual SA examination was done by Donglin et al. [67]. In this, a review is introduced which characterizes the various strategies utilized for the visual assessment examination. In this sort of investigation, pictures are utilized to foresee the conclusions of the individual. The overview is fundamentally centered around the leading-edge techniques that are utilized in the picture examination process. This review portrays the new stage for a specialist because principally research is done on the content yet visual sentiment ontology is another idea for accomplishing something else.

It was proposed by Tan et al. [144] that the Sentiment examination performs the better process of decision making provided to a specific individual item or any service. This work predicts the extremity of words and after that orders them into positive and negative emotions with the point of recognizing opinions and attitudes that are communicated in any language or structure. This technique incorporates different steps of preprocessing before sustaining the content to the classifier and the Map-Reduce algorithm will be utilized for getting a precise choice about items from various views of input from clients.

A deliberated study about help arranged circulated computational framework dependent on the Cloud ideas called Distributed Computational Service Cloud was conducted by Han et al. [54]. This sort of Cloud has versatile services of the Grid, which are executed with Web Services-Resource-Framework-consistent. In this choice, the services of the tree are utilized, which gives a computationally serious information mining algorithm.

Further, the exponential development of the stored and processed data on the Internet was presented by Hamza et al. [131]. It has turned out to be more than it should be expected,to utilize distributed computing advancements and significant information extraction systems (information mining) to lessen the processed and stored data volume. It prompts to decrease the material expenses. In this, a new cloud computing design dependent on another SVM procedure is created for producing learning models that permit data filtering to extricate the most applicable information.

Also, it was observed by Ikram et al. [61] that computational biological systems, several biomedical and computational bioinformatics applications are developing at a fast pace with an expanding interest in power processing. The approach based on grid clusters was effective however it presented the need for systems. In this novel cloud computing-based neural system a structure is presented and aftereffects of MSA i.e. Multiple Sequence Alignment algorithms in the architecture of the cloud are observed.

Also, a study by Kim et al. [77] portrayed that with the predominance of cloud computing, query privacy, and data privacy from enemies, databases should be scrambled before being redistributed to the cloud. In any case, the classification scheme of the KNN plan exists over the encoded databases in the cloud. Since the current plan experiences high calculation overhead, a protected and effective KNN classification calculation that hides the subsequent class label and information access examples are presented. Likewise, this strategy can bolster proficient KNN order by utilizing the encoded list and Yao's distorted circuit.

A proposal about the novel process by joining classifiers for the development of individual classifiers' performance was given by Ikonomakis et al. [60]. Numerous approaches have been proposed for the formation of the gathering of classifiers including a diverse subset of trained data with a solitary learning technique and preparing parameters with a solitary preparing strategy.

It was suggested by Kang et al. [74] that the global features are class-independent whereas the local features are dependent on the features. It also presented that the local dictionaries are known to be class-independent. The best technique based on text categorization is acquired by utilizing a blend of both local dictionaries and local features.

A study by Isa et al. [63] revealed a new order approach utilizing the SVM classifier at the back end and naive Bayes method at the front end to characterize the records to the correct classification. This hybrid of SVM classifier and Naive Bayes vectorizer improved classification-based accuracy contrasted with a technique named Pure Naive Bayes classification.

A few distinct techniques for consolidating numerous classifiers for text-based categorization were investigated by Bao et al. [12]. Numerous scientists have demonstrated that consolidating distinct classifiers can improve the accuracy of classification. It is seen from the analysis of the outcome between the individual classifier and the combined method and found that the consolidated technique exhibition is superior.

A novel hybrid technique was proposed by Cheng et al. [44] wherein the LPEBP i.e. Learning Phase Evaluation Back Propagation Neural Network is acquainted with improving the customary BPNN. A strategy to diminish the measurement and build the latent semantics was generally proposed. It demonstrates that the LPEBP is a lot quicker than the BPNN, subsequently improves the performance of conventional BPNN.

Kaminskas et al. [73] put light on Rocchio's Algorithm, which represents a vector space technique for filtering or document routing in retrieving information. It manufactures a prototype vector for each class utilizing a preparation set of documents and figure comparability between test record and every one of the model vectors, which allot test documents to the class with the most extreme closeness.

In another study, the exchange of views was done by Brücher et al. [20], that Naive Bayes classifier is a straightforward probabilistic classifier-based approach. A progressively elucidating definition for the probability-based model would be an autonomous feature model. These suppositions make unessential feature order and demonstrate that the presence of one feature or component doesn't affect different features. Subsequently, the calculation of the Bayesian order method is increasingly proficient. The Bayes classifiers can be trained precisely by requiring a relatively modest quantity of trained information to appraise the parameters for classification.

Gitanjali et al.[46] proposed a simple approach to Text Classification, which focused mainly on text mining and processing of natural language. The findings showed that the proposed CNN-Logistic regression approach substantially improves accuracy due to the improvement of feature patterns.

The confusion between two models which have Naive Bayes's assumption was cleared by McCallum Andrew et al [97]. It described the differences and details of both models and also compared both models on five text corpora. It founded that multi-variate Bernoulli performed better with small sizes of words and multinational performs better with large word sizes and reduced the error by 27% over the multi-variate model with any size of words.

The maximum entropy-based techniques for classification of text were proposed by Nigam Kamal et al [109]. It was a probability distribution technique that was used for many tasks like the modelling of language, segmentation of tasks, etc. The principle of maximum entropy was that when nothing was given then distribution must be uniform. It has a unique solution which was found by an enhanced iterative scaling algorithm.

Nigam Kamal et al [110] suggested an algorithm that helps to learn the labelled and unlabelled documents based on Expectation-Maximization (EM) and a naive classifier of Bayes. In it, they firstly trained a classifier with the labelled documents and probably labelled the unlabelled documents. After that, it trained a new classifier with the help of labels of all the documents and iterates to merging.

Moreover, another algorithm was given by Tong Simon et al [147] in which support vector machines were used for active learning. With the parameter space and feature space, three algorithms were found which decreased the version space. It showed good performance in both inductive and transductive settings and also it was seen that it also decreased the requirement of labelled training instances.

An approach was suggested by Lodhi Huma et al [91], which helps in the classification of text documents based on a specific kernel. It mainly focused on the classification of text based on Support Vector Machines. This kernel was used with other kernel-based learning systems in clustering, ranking, categorization, etc.

An experimental comparison of 12 methods for the selection of features was initiated by Forman George [42]. It was evaluated on the reference point of 229 text classification problem details which were collected from Reuters, TREC, OHSUMED, etc. In it, results were calculated on the outlook of accuracy, precision, etc. From the results, it was seen that a new method of feature selection called 'Bi-Normal Separation' showed much better results than other techniques in some of the situations.

A study was conducted by Zhang Xiang et al [160] on the character level convolutional networks for text classification. The analysis showed that character- level ConvNet was an effective method. A model performes well depending upon various factors like dataset size, whether the texts were curated, and the choice of the alphabet.

A recurrent convolutional neural network for the classification of text was introduced by Lai Siwei et al [81]. It also engaged the max-pooling layer which itself judges the words that played an important role in the categorization of the text so that components can be captured. The results revealed that the method shows better results on many datasets especially on document level datasets.

A simple and efficient baseline for text classification was intended by Joulin Armand et al [70]. In it, the features of words were combined to make a good representation of sentences. During many tasks, the fast text gets better results with the proposed baseline and becomes much faster. Deep neural networks had much better representational power than other shallow models.

A semi-supervised bug triage approach based on an NB classifier with EM was propounded by Xuan Jifeng et al [158]. It also enhanced the accuracy of classification with labelled and unlabelled bug reports up to 6%.

2.2 Review of Sentiment Analysis Techniques

Research on sentiment analysis was started in early 1990. The term sentiment analysis along with opinion mining was first introduced during the year 2003. The research on this analysis was limited only to the identification of subjective content, sentiment adjectives, and interpretation of metaphors.An algorithm was presented by Weibe J.M.[156], in which subjective characters were identified in fictional narrative text based regularities in the text.

An intelligent idea of text-based systems to refine the information access task was given by Hearst M. A. [57], while extensive examinations were carried out by Weibe J.M. [156] to find out if the naturally occurring narratives and regularities came up with an algorithm that would track the point of view based on these regularities. Early work on the sentiment-based classification of whole archives has regularly included either the utilization of models roused by intellectual phonetics, as mentioned by Hearst M. A. [156] or the manual and semi-manual development of discriminant word lexicons [57].

A philosophy was exhibited by the great researchers Das and Chen [33] for constant sentiment extraction in the space of finance; working with messages from electronic stock message sheets, endeavour to consequently name each such message as a "buy", "sell" or "neutral' proposal. This classifier accomplishes a precision of 62 % (the upper bound, human understanding rate, was 72%). The key features of this technique were the manual compilation of a discriminant-word lexicon and the marking of terms from a few thousand texts.

The analysis exhibits more modern elements based on phonetic information or heuristics, combined with different element choice and smoothing strategies. Contingent upon the particular trial setting, results are blended with applying standard machine strategy to straightforward unigram highlights. For example, execution can be enhanced by trigrams and bigrams under certain settings. Still, a fascinating component in the light of reliance parsers neglected to make strides execution on the test sets. Consequently, different sorts of components on the other hand included choice plans that have been investigated by the researchers [43, 96].

One fascinating issue emerging from space particular nature of this issue about directed learning is how to adjust the classifier to another area. This issue was addressed by the researcher's Dave, Lawrence, and Pennock [34] where "follow-on" errand was the classifier prepared on the pre-collected dataset. Diverse ways were investigated by Aue and Gamon [43] to deal with the redo of a sentiment grouping framework to another objective space without a lot of marked information. The distinctive sorts of information considered, range from protracted film audits to short-express level client criticism to web reviews.

Due to the critical contrasts in these areas, the distinctive topic, and besides diverse styles and lengths of composing, just applying the classifier took in on information from one area can scarcely beat the pattern for another space. Additionally, an option wellspring of marked information: emoticons utilized as a part of Usenet newsgroup postings was investigated by the author. The thought was to concentrate messages around grin or scowl emoticons and guide the emoticons into polarity marks.

Rather than supervised or semi-supervised learning, a shot was taken by Turney in his work in 2002 [149] at a grouping of audits connected to a particular unsupervised learning system. Data was processed utilizing measurements on information assembled by a web crawler. The author worked with surveys on distinctive subjects, reporting correctnesses running more than 64% (film audits) to 84% (car audits). Different avenues were explored by Beineke [16] regarding an expansion of Turney's strategy that likewise used a small arrangement of marked information. In general, unsupervised ways of dealing with archive-level polarity grouping include, as is typically indicated in the writing, programmed naming of terms or phrases by their sentiment polarity or " semantic introduction " Few reviews are completely about building such lexicons. An early approach in light of phonetic heuristics was introduced by Hatzivassiloglou [17]. The primary thought is to utilize data removed from conjunctions between descriptive words in an extensive corpus, for example, in the event, where two descriptors are being connected by at the same time, this proposes that they are of inverse introductions; alternately, being connected by and could be a proof for the descriptive words having a similar introduction. The assignment is then cast into a bunching issue with imperatives given these heuristics. Later reviews in this bearing begin with a small arrangement of seed words, at times, only two words with inverse twelve polarities, and utilize these seed words to (once in a while incrementally) tag different words. Corpus-based methodologies look at co-events with seed words in the light of extensive accumulations of content [17] or look for the setting subordinate names by considering nearby imperatives. On the other hand, abusing learning encoded in WordNet, for example, relations like synonymy, antonym, and hyponymy have been investigated by individuals, Kim [58, 78, 95], as mentioned by the researcher.

A model was suggested by Pak and Paroubek (2010)[111] for the objective, positive and negative classification of tweets. A Twitter corpus using Twitter API to collect tweets and automatically annotate those tweets using emoticons was created by them. A sentiment classifier based on the Naive Bayes multinomial algorithm, using features such as N-gram and POS tags, was designed using that corpus.

Two models were introduced, a Maximum Entropy and NB Bigram model and a for tweet classification by Parikh and Movassate (2009)

[112]. It was found that classifiers of the Naive Bayes performed much better than model Maximum Entropy.

A sentiment analysis approach was introduced by Go and L.Huang (2009) [47] for Twitter data using a remote control, in which their training data consisted of emoticon tweets that acted as noisy labels.They had unigrams, bigrams, and POS as their feature space. The conclusion drawn was that SVM outperformed other models and was more anagrammatic.

An automatic two-phase sentiment analysis method for the classification of tweets was developed by Barbosa duvet al. (2010) [124]. Tweets were categorized as objective or subjective and instead, the subjective tweets were further categorized as positive or negative in the second step.

Twitter streaming data generated by the Firehouse API, was used by Bifet and Frank(2010) [19] which included all real-time publicly accessible messages from any user. Stochastic gradient descent, Multinomial naive Bayes, and the Hoeffding tree were experimented with. They concluded that when used with a reasonable learning rate, the SGD-based model was better than the other models.

A 3-way model was developed by Medhat et al. (2011)[98] to classify sentiment into neutral, negative, and positive classes. They experimented with models such as a model based on unigrams, a model based on functionality, and a model based on tree kernels. Tweets were described as a tree-based model for tree kernel. 100 features are used in app oriented model, and over 10,000 features are included in the unigram model.They concluded that characteristics that combine the polarity of the previous word with their part-of-speech (pos) tags are most important and play a major role in the classification task.

An approach was proposed by Davidov et al.,(2010) [35] for using Twitter user-defined updates in tweets as a classification of the feeling type. The KNN strategy gives sentiment labels by creating a characteristic vector in the training and test set for each case.

2.2.1 Opinion mining and Summarization

There may be dull sections of an opinion-bearing archive that may not be keen on the readers. According to the authors Hurst and Nigam, one likely way to deal with focused sentiment polarity on a given point is to first apply a sentence-level classifier to identify topical sentences. [59]. It may be beneficial to condense the views assembled by the particular angles tended for surveys that cover a few sections of the subject, and there has been work that explored recognizing characteristics and opinions related to these highlights from product surveys as described in[157].

2.2.2 Perspectives and Viewpoints

Some early work on non-verifiable based content analysis managed points of view and perspectives [157]. The Multi-Perspective Question Answering is a system that extends the concentrates on address noting errands that require the capacity to break down opinions in content so that responses to inquiries, for low-level assessment explanations were produced and assessed. This encouraged the investigation of various fascinating issues, for example, distinguishing supposition holders and breaking down opinions at expression level as mentioned by the authors in [154].

Influences and feelings of people have considered different influence sorts [87, 88] and additionally, computational methodologies for computational approaches for humour recognition and generation says the authors Michala and additionally being investigated with regards to casual content assets, for example, weblogs. In any case, clients of computer-mediated correspondence have discovered their methods for overcoming the absence of individual contact by utilizing emoticons.

The main emoticon was utilized on September 19, 1982 by educator Scott Fahlman in a message on the software engineering notice leading board of Carnegie Mellon University. In the message, Fahlman proposed to utilize the character arrangements-)" and - (" keeping in mind the end goal to unmistakably recognize jokes from more genuine matters. Individuals began sending yells, hugs and kisses by utilizing graphical images shaped by characters found on the keyboard. After ten years, emoticons had discovered their way into ordinary computer mediated correspondence and had turned into the paralanguage of the web, explains the author Marvin in 1995. By then, 6% of the messages on electronic mailing records as mentioned by Rezabek and Cochenour in 1998 and 13% of UseNet newsgroup posts were assessed to contain emoticons.

Consequently, nonverbal cues have developed in computer communication. It ought to however be noticed that these nonverbal cues in computer communications are unique about nonverbal cues in face-to-face correspondence. Genuine cues like laughing and sobbing are frequently thought to be automatic methods for conveying everything that needs to be conveyed in face- to- face correspondence purposefully, explains the author Kendon in 1987. All things considered, emoticons empower individuals to show unobtrusive inclination changes, to flag incongruity, mockery, and jokes, and to express, accentuate or disambiguate sentiment.

In any case, the emoticons have as of now been abused to a restricted degree, for the most part for computerized information comment. For occasion, a rough refinement between a modest bunch of positive and negative emotions has been utilized as a part of the request to naturally produces information sets with positive and negative examples of content.

Researchers, Sudharshan M., Prabhu J., Saravanan M., and Prasad G. [118] discussed the use of the Rapid clustering method to analyze the characteristics in a social network and to cluster dataset. The authors proposed an innovative clustering technique named the Rapid Clustering Method (RCM), with the use of Subtractive Clustering. Here, Xin Chen, Krishna Madhavan, and Mihaela Vorvoreanu [27] use Social Web Analysis Buddy (SWAB) to analyze student-posted content on social media sites to facilitate the understanding of human behaviours and social tendencies. This technique consists of qualitative analysis and large-scale data mining which will allow researchers to build modelling algorithms. The main idea here is to demonstrate that it is possible to analyze data on tweeter through Cloud infrastructure.

Javier Conejero, Peter Burnap, Omer Rana, and Jeffrey Morgan [30], used the COSMOS (Collaborative Online Social Media Observatory) platform to support sentiment and tension analysis on Twitter data, as well as demonstrating how this platform can be used with Map in the OpenNebula Cloud environment. Federico Neri, Carlo Aliprandi, Federico Capeci, Montserrat Cuadros, and Tomas [108] described a study on sentiment analysis done on more than 1000 Facebook posts. The study of H. K. Chan, E. Lacka, R. W. Y. Yee, M. K. Lim [24] shows that social media data can be used by the exploitation of comments and statistical cluster analysis to identify the interrelationships among important factors.

David Ediger Karl Jiang [40], presented GraphCT, a Graph Characterization Toolkit for massive graphs representing social network data. By using GraphCT, actors can be ranked within Twitter messages and help analysis to focus on smaller data subsets.

Haruna Isah, Paul Trundle, Daniel Neagu [64] report that the development of a framework is focused on the compilation and analysis through machine learning, text mining, and sentiment analysis of the views and perceptions of consumers of drugs and cosmetic products.

Victor Joo Chuan Tong, Zhaoxia Wang, and David Chan [155] wanted to address the shortcomings of existing approaches by not only enhancing the algorithm's precision, but also by demonstrating the possibility of conducting non-English language analysis. The authors performed a survey on the current sentiment analysis research literature and explored numerous drawbacks of the existing analytical methodology and suggested a technique using a modern sentiment analysis technique.

Amir Hossein Akhavan Rahnama [120] perform real-time sentiment analysis.This research work demonstrates that it is difficult to store data instances with real-time data processing and can thus use online analytical algorithms. This provides better results compared to uni-processor classifiers in terms of both precision and performance enhancement.

Yafeng Lu, Feng Wang, Steffen Koch, Robert Kruger, Dennis Thom, Thomas Ertl, and Ross Maciejewski [93] used predictive analytical methods to explore how to predict in such a framework results from a user study.This work incorporated comparisons of similarity and model cross-validation of the selection mechanisms to assist analysts through model building and prediction.

Bing Liu [90] introduces the field and presented some technical challenges that turn around sentiment analysis into a multi-faceted problem containing subproblems. Also discussed opinion spam and the issues related to it. The author illustrates the history and potential of research-based sentiment analysis over the past few years. Jingwei Liu, Erick Blasch, Yu Chen, and Genshe Chen [89] demonstrate that it is possible to scale up to analyze the sentiment on millions of movie reviews with increasing throughput. To do that, the evolution of the Naïve Bayes Classifier in a large collection of related sets of information was evaluated.

Akshay Amolik, Niketan Jivane, Mahavir Bhandari, M. Venkatesan [116], with the help of classifiers such as Naive Bayes and support vector correctly classified these tweets as neutral, negative, and positive.

Lina L.Dhande, Girish K. Patnaik [36] had used Naïve Bayes classifier by combining the neural network on a standard dataset that improves the accuracy and performance of sentiment classification on positive or negative user reviews.

Reshma Bhonde, Binita Bhagwat, Sayali Ingulkar, and Apeksha Pande [18]discussed Sentiment-based text analysis methods that automatically identify the polarity of text that can assist with more refined analysis of these circumstances. A graphical analysis that is easier, in this case, was used to solve these situations.

Neethu M. S. and Rajasree R. [107] try to analyze Twitter posts about electronic products like mobiles, laptops, and tablets, by using a knowledge base approach and machine learning approach. The effect of domain information through sentiment classification was identified.

Changbo Wang, Zhao Xiao, Yuhua Liu, YanruXu, Aoying Zhou, and Kang Zhang [153] present SentiView, an interactive visualization framework aimed at analyzing public feelings on common Internet topics.

Bin Lu, MyleOtt, Claire Cardie, and Benjamin Tsou K. [92] investigate the efficiency of topic model-based approaches to two multi-aspect of sentiment analysis tasks:

- Multi-Aspect Sentence Labelling
- Multi-Aspect Rating Prediction.

Songbo Tan, Xueqi Cheng, Yuefen Wang and Hongbo Xu [143] used Adapted Naïve Bayes (ANB) to improve performance of the Naïve Bayes Transfer Classifier. This new method allowed every researcher to gain knowledge from new domain data based on sentiment analysis demand.

Owen Rambow, Apoorv Agarwal, Ilia Vovsha, Boyi Xie and Rebecca Passonneau [4] talked about POS-specific prior polarity features in the first part. Secondly, to avoid the need for tedious function engineering and tree kernel to perform the same level exceeding the state-of-the-art baseline, they explored the use of a tree kernel. Alexander Pak and Patrick Paroubek [114] developed a subject on microblogging. They used both opinion mining and sentiment analysis to perform linguistic analysis and explain the phenomena. This work classified sentiments that were neutral, negative, or positive on a document.

Taboada [141] explores methodologies to classify sentiment on Twitter data. Here the author used lexicon-based methods, Simple Word

Count, and Feature Scoring approach to compare ten classifiers on existing Twitter data. The author compared two sets of features that are Bag-of-Words with N-Gram and Part-of-Speech linguistic annotation.

Luciano Barbosa and Junlan Feng [14] show that labels with more noises were provided through a few sentiment detections on websites over Twitter data. A proposal to solve this problem is by detecting sentiments on Twitter messages or tweets by exploring some characteristics of how tweets are written and the meta-information is used.

Ming Hao, Christian Rohrdantz and Halldór Janetzko [55] showed high data volume pixel cell-based sentiment calendars and high-density geo maps to virtualize huge data volumes in a single view.

Hassan Saif, Yulan, and Harith Alani [129] used semantics to sentiment analysis methodology to predict sentiment for three different Twitter datasets. The work also makes the comparison against an approach based on sentiment-bearing topic analysis to find this semantic through the classification of negative and positive sentiments.

Tan [142] showed that clients with common comparative opinions were probably to be connected. The creators proposed the model that was produced either by taking after the network that has been made by labelling diverse client Twitter adherent or follower. The creators clarified that by employing the data connection of Twitter there will be a change in client-level sentiment analysis.

Chen [26] utilized the feed-forward BPN system and sentiment orientation to figure the outcomes at every neuron. The creators proposed an approach based on the neural arrangement. The proposed procedure is a blend of machine learning classifiers and semantic introduction records. Keeping in mind the end goal to acquire proficiency in methodology, semantic introduction lists utilized as contributions for the neural system. The proposed methodology beats other neural systems and conventional methodologies by increasing efficiency in both preparation and also classification time.

The authors Anton and Andrey evaluated the current systems and built up a model for automatic sentiment analysis of Twitter messages utilizing unigram, bigram, and together called a hybrid model. Kouloumpis [79] investigated the upside of semantic elements for analyzing the sentiment of messages of Twitter. The authors research the components that collect knowledge about natural and casual dialects that are utilized as a part of microblogging also as the advantage of existing lexical assets. The authors utilized the supervised learning method to the issue and to gather it hashtags are utilized. The authors inferred that in the experiment part-of-speech feature is worse for sentiment analysis when it comes to the area of microblogging on Twitter and it is affirmed that for gathering data, hashtags are extremely valuable so that messages with negative and positive emoticons are classified.

Nasukawa and Yi [106] proposed an approach for sentiment analysis to discover sentiments connected with negative or positive polarities from a record for a particular subject, rather than arranging the entire archive into negative or positive. The real issues in sentiment analysis are the opinion about the focus of the announcement which has negative or positive behaviour towards the subject and discovery of how sentiments are portrayed in writings. The authors stated that it is fundamental to unmistakably discover the semantic connections between the subject and the sentiment expressions to build precision for the analysis of sentiment. To recognize the sentiments in news articles and website pages, their proposed framework got high accuracy of 75-95%.

Gautami Tripathi[148] suggested a model for sentiment analysis of various movie reviews using a combination of natural language processing and machine learning approaches.

Sreeja Rajesh [121] has done a detailed study on how the administrations given by Social media is useful for understudies in Education. This paper likewise tries to clarify the adequacy of Social Media in youngsters. The creators expressed that the web-based social networking administrations and assets can make use by the clients and this paper did the investigation of its adequacy in training recorded and discovered, that it is an exceptionally helpful tool in instruction purposes and past.

Alessia D'Andrea [31] gave a review of the diverse sentiment classification approaches and devices utilized for sentiment analysis. Likewise gives the distinctive arrangement systems its favourable circumstances and confinements, diverse instruments for various procedures utilized as a part of sentiment analysis. Various application areas of the use of sentiment analysis are also explored in the paper, such as industry, politics, public behaviour, and finance.

Zhunchen Luo [94] discussed the problem of finding opinionated tweets about a given topic. They automatically construct opinionated lexical from sets of tweets matching specific patterns indicative of opinionated messages. It also stated that to improve query-dependent opinion, retrieval of topic-related specific structured tweet sets can help.

Kalarani P. [71] discussed an overview of research challenges and application areas in opinion mining and the techniques and tools used for opinion mining. They have discussed various research scopes and the architecture of opinion mining.

Nidhi Mishra, Jha C.K. [100] in this research paper they focused on the classification of opinion mining techniques, which conveys the opinions of users, positive or negative, at different levels. The essential method of predicting views allowed them to extract feelings from the web and forecast customers' online tastes, which could prove useful for marketing research.

An Opinion miner has been developed in this work for mining the opinions of customers and extract detailed product entities on which reviewers express their opinions about the product. They have described the architecture and main components of the system. And the shortcomings were the long sentence comments which were not analysed and it brings the system performance down. When used with pronoun resolution it produced a more false positive.

2.2.3 Sentimental Analysis in Social Network

Generally, massive quantities of statistics are collected from social networks, blogs and other media and are subtle to the huge web market. This vast record consists of very significant data related to views that can be used to support agencies and various business and medical industry components. It is not always feasible to manually track and extract these useful documents, so sentiment analysis is needed.The assessment of feelings is the phenomenon of extracting feelings or feedback from evaluations shared by users about a particular online issue, place or product. It clubs the emotions into classes like "high quality" or "terrible". Thus concerning the subject in context, it defines the general attitude of the speaker or a writer. The following categories could be used to explain social media platforms:

Relationship Networks: The most common type of social media that are used these days allows the users to either communicate privately or publicly with their entire network. It keeps varying according to need, for example, a professional network like LinkedIn is used for finding jobs, and updating professional information, it is also used to connect to other professionals around the globe. Facebook is an informal way of connecting with people. Twitter is used for tweeting short messages over the network.

Media Sharing Networks: Media sharing networks could be categorized according to the media that is allowed to share, for example, video-sharing social networks like Vimeo and YouTube, picture sharing sites like Instagram and Pinterest. Users are allowed to connect to other accounts; send, share, and view the media posted by others.

Online reviews platform: With the growing popularity of websites like Amazon.com, Flipkart.com, and many others, opinions and ratings are given by people across the globe on different products. With the help of reviews, comments, and ratings the users make the right choice. Geo-location tagging has been used for a better recommendation system. Trip advisor is a common example of an online reviewing network.

Social publishing platforms: Social publishing platforms are used for publishing content socially. Examples of social publishing platforms include Tumblr and WordPress.

Discussion forums: These forums are used to share knowledge to help others having a similar query or problem. Stack overflow and Quora play a major role as discussion forums.

E-commerce platforms: Small businesses or individual entrepreneurs sell their products over this type of network. Major examples would be Flipkart and Amazon. Over the years, Twitter and Facebook are also expanding themselves to e-commerce.

Bookmarking sites: Stumbleupon, Pinterest, and Flipboard are the sites used to record and save data from different sources on the internet. It can also be shared with others.

Interest- based networks: Music-based websites like Bruno Mars, Kid Rock are used by musicians and music lovers.

2.2.4 Components of Sentiment Analysis

Sentiment analysis is usually performed in a sequence of sub-tasks. The given text is first broken down into different components. These component identifications are performed in a sequence of steps. Different components included in sentiment analysis are holder, aspect, and polarity. A specific task is performed by every component in the system.

- Holder denotes the entity that holds the sentiment. Opinion holders are also considered vital in the analysis. For example, they provide clear views on a circumstance, place, company, or individual who expressed their disagreement with the purchased house in the news analysis. It is the opinion holder which is published in news articles.

- Polarity is the property of sentiment. Sentiment analysis is based on this polarity identification in the given text. This could classify the given text as good or bad, accept or reject, positive or negative, yes or no, and any of this sort. Polarity can be two-folds or three-folds, classification of the text for the third property neutral. When there are opinions about the text, it is classified into the third property neutral.

- The particular part or feature of the target that the sentiment is expressed is known as aspect. An attribute or part of an object is a trait or feature. For example, a fruit's colour, the size of the insect, or a car's consumption of petrol.

2.2.5 Classification Process of Sentiment Analysis

Feature selection is most important in any sentiment classification. The parameter or the opinion about a product is considered for sentiment analysis. For example, if the sentiment polarity is to be analysed for a mobile phone, then the specification of the mobile phone like camera, screen width, memory, and the price is taken into consideration. These parameters are the features to be considered for finding the sentiment. There are 3 main levels of classification in sentiment analysis:

Document Level

Sentiment Analysis performed at the document level aims at classifying an opinion document to express positivity, negativity, and neutral in the sentences present in the document. The whole document is considered when it comes to finding sentiments on a particular topic. The principal contribution is the Bag-of-Opinions (BoO), which predicts document-level polarity and intensity of online surveys.

Sentence Level

Sentiment Analysis performed at the Sentence level aims at classifying sentiments from each sentence.The approach would first require deciding whether the sentence is subjective or factual. If it is found to be subjective, then it is possible to use sentence-level sentiment analysis to find the sentence's polarity. In most cases, sentiments are not necessarily subjective. One cannot find much difference between sentence levels and document level, because the collection of sentences forms a document. It does not provide much information on all things that are required to find feelings when the classification is performed at the text level or sentence level.

Aspect level

Sentiment Analysis performed at the aspect level aims at classifying the sentiments concerning the specific aspect of entities. The process would involve identifying the entities and their aspects. Then the opinion holders can give a different opinion of the same entity like this sentence 'the sound quality of the speaker is not that good, but battery life is amazing'. The dataset plays a major role in this field. In this case, our major source of dataset is from Twitter. Twitter is a social networking site that also provides news services wherein the registered users are allowed to post and read 140 character messages called tweets. Registered users can control the viewers of the tweets so that non-registered users can also read them. It was created in March 2006 by Jack Dorsey, Evan Williams, Biz Stone, and Noah Glass.

It has gained a lot of shares and grown rapidly over the years gaining worldwide popularity with more than one hundred and fifty million users posting around five hundred million tweets a day. Whatever is posted by the users on Twitter, acts as reviews, that are important and help in taking decisions. In addition, sentiment analysis should not only be extended to Twitter tweets, but also to financial markets, news stories, and political debates. If the political discussion in the US presidential election were to be considered, it would be easy to find out people's opinions on a specific candidate. The election results can also be predicted. Not only Twitter, but other social networking websites are also a very good source of information because these are places where people post, share and discuss topics freely; making it a very good source of data. Over the years, a lot of research work has been going on sentiment analysis. A lot of algorithms were also proposed. In this research work, a comparison is done of the normal Naïve Bayes approach, using unigram and bigram.

2.2.6 Cross Domain Sentiment Classification

Automatic classification of sentiment is imperative for various applications, for example, assessment mining, supposition outline, relevant publicizing, and market examination. Normally, sentiment classification has been demonstrated as the issue of preparing a binary classifier utilizing audits commented for positive or negative sentiment. Nonetheless, the assumption is communicated distinctively in various areas, and commenting on corpora for each conceivable space of intrigue is expensive. The sentiment is orthogonal to point and opinion order is more troublesome than subject classification. Generally, they demonstrated that Sentiment investigation is an area of a particular issue, and it is difficult to make domain-independent classifiers.

One conceivable way to deal with cross-area order is to prepare a classifier on a domain-mixed data set instead of one specific domain. The diverse sorts of information are considered from extensive motion picture audit to short and state-level client input to Web studies. Due to the huge contrasts in these spaces (distinctive topic and also extraordinary styles and lengths of composing), applying the classifier took in the information from one space can scarcely beat the benchmark for another area.

Recursive Deep Models for Semantic Compositionality

Semantic word spaces have been extremely helpful be that as it may; they cannot express the significance of longer expressions principally. Advance towards understanding compositionality in errands, for example, estimation identification requires wealthier managed preparation and assessment assets also need more intense models of creation. Semantic word spaces have been extremely helpful however cannot express the significance of longer expressions principally.To solve this, a paper was presented by Richard Socher et al on Sentiment Treebank [136]. In the parse trees of 11,855 sentences, it integrates fine grained supposition names for 2, 15,154 expressions and reveals new difficulties for opinion compositionality. They introduced the Recursive Neural Tensor Network to answer them.

At the point when the researchers prepared the new Treebank algorithm, this model outflanks every single past strategy on a few measurements. It calculates the best in class, in a single sentence positive or negative grouping from 80 % to 85.4% is achieved. The precision of anticipating fine-grained assessment marks for all expressions achieves 80.7%, a change of 9.7 % over pack of elements baselines. Finally, for both positive and negative expressions, it is the only model that can reliably capture the results of nullification and its extension at distinct tree levels. For all decision tree sentiment analysis or opinion mining, this served as the fundamental model.

2.2.7 Features of Sentiment Analysis

Features usually used in sentiment analysis are also included, including characteristics reflecting information from a sentiment lexicon and features of Parts Of Speech (POS). In this job, functions are also introduced to capture some of the more domain-specific language of micro blogging. In text mining and natural language processing activities, n-gram characteristics of texts are primarily used. Different sets of characteristics have been used in this analysis for the ion classification experiment. Unigrams or bigrams are used for the baseline.

Lexicon features: Lexicon based approach is used to extract sentiments or opinions from the text. There are generally two techniques for sentiment analysis. One is using a supervised learning algorithm and the other can be dealt with as unsupervised learning. One uses an algorithm for supervised learning and the other can be treated as unsupervised learning. In general, a supervised learning algorithm assembles a classification display on a significant annotated corpus. The consistency is mainly provided by the standard of the annotation, and the training method will take a long time for the most part. Other than that, when the algorithm is applied to another domain, the outcome is normally not great. Contrasted with supervised learning, lexicon-based unsupervised learning utilizes a sentiment dictionary, which does not require storing a substantial information corpus and training, which makes the entire procedure much quicker.

- **TF-DF** -Term frequency or Identifier frequency, speaks to just the number of the terms or "words" being taken into account. These terms are possibly unigrams, bigrams, or higher request n-grams. Which ones of these yield better outcomes is still not known. Author of [130] assert the predominance of unigrams over bigrams in a motion picture survey sentiment analysis, though the author of [65] contend bi-grams and tri- grams give better outcomes on the premise of their product review classification analysis.

- **POS - Part Of Speech Tags**:English is, by default, a questionable language. There may be more than one synonyms of one particular term that depend on its usage meaning. POS is used to disambiguate context, which is thus used to monitor the choice of function [48]. These marks, for instance, can be used to distinguish adjectives and adverbs since [145] is typically used as sentiment pointers. Later on, they realised that the execution of modifiers on the assumption of recurrence was more regrettable than the same amount of unigrams chosen.

- **Syntax and negation**:-To boost execution, the use of collocations and other syntactic elements can be used. Algorithms using syntactic components and algorithms using n-gram elements were counted to achieve a similar execution in the order of short messages.

Twitter Specific Approaches: The fundamental distinction between sentiment analysis of Twitter and reports is that Twitter based methodologies are more particular towards deciding the polarity of words, basically descriptive words. Though the report based methodologies are particular towards the undertaking of deciding components in the content.

There are main three approaches for Twitter sentiment analysis

- Lexical Analysis Approach

- Machine Learning Approach
- Hybrid Approach

Utilizing one, or a blend of the distinctive methodologies, one can utilize one or more lexical and machine learning procedures. In particular, one can be a combination of lexical and machine learning techniques. Particularly one can utilize unsupervised procedures, supervised techniques or a blend of them. One can start the process with auditing the lexical methodologies which concentrate on building fruitful word dictionaries, then the machine learning approaches which are fundamentally concerned with highlight vectors, and lastly a combination of both, a hybrid approach.

Lexical Analysis Approach: This approach usually uses pre-labelled words in a lexicon or lexicon. In comparison to the lexicon, any word available in the material is considered. When a word is present in the reference word, its polarity esteem is added to the total polarity rating of the content. For example, if a match has been marked with the word "excellent," which is explained as positive in the word comparison, then the aggregate polarity score of the blog is increased. If the content's aggregate polarity score is certain, then that content is called positive, otherwise, it is named negative [37].

Machine Learning Approach: The other primary avenue of research inside this range has used managed machine learning strategies. Within the machine learning approach, a progression of highlight vectors is picked and a gathering of labelled corpora is accommodated preparing a classifier, which can then be connected to an untagged corpus of content. In a machine learning approach, the choice of components is critical to the achievement rate of the classification. Most generally, as function vectors, a set of unigrams, single words from a document, or n-grams, are chosen as at least two words from documentation in consecutive requests. The number of positive terms, the number of negative words, and the length of the study are other highlights. Since a greater part of sentiment analysis approaches utilizes machine learning strategies, the elements of content are frequently spoken to as highlight vectors. The accompanying is highlights utilized as a part of sentiment analysis.

Hybrid approach: Some methodologies utilize a mix of different methodologies. One combined approach is taken from making a combination of KNN and SVM. The second combined approach is taken from making a combination of CNN and LR. This approach begins with two-word lexicons and unlabeled information. It generated a pseudo-report containing each of the expressions of the chosen lexicon with the two prejudicial word lexicons, negative and positive. After that, the resemblance of the cosine between these pseudo-records and the unmarked reports is noticed. For cosine similarity, either negative or positive sentiment is assigned to a record.

2.3 Review of Encryption Techniques

Almorsy et al.[6] addressed a cloud security management framework that is based on the FISMA standard which requires users and cloud providers to be accredited for security. It strengthens collaboration between cloud providers and service users to monitor security. The NET framework and SaaS application security management are introduced using that approach.

Yibin et al.[83] implemented a smart cryptographic solution that prevents partial data from being accessed by the cloud provider. This approach breaks the file into sub-files and stores it on cloud storage those files. Another solution to determining when to break data packets to minimize running time is also proposed. The suggested solution offers protection and reliability with better computing time.

Various cryptographic algorithms were proposed by Diwan et al. [38], which were compared and taken into account to ensure confidentiality of data. In these various cryptographic algorithms, various parameters are compared, including block size, type of key length, and characteristics. The concept for another cryptographic algorithm that can be used to ensure cloud data protection was received by the author.

Sood et al.[138] addressed a combined solution that offers cloud storage data protection. In this work, different techniques are combined to provide the sender with the ends of the receiver with efficient protection. The protection of data given to the user is focused on data confidentiality, integrity, and availability. The protected socket layer provides data protection using an encryption mechanism, and Media Access Control provides integrity. With all the users using the Login Id and Password operation, the security is improved.

A Cloud computing security framework was addressed using cryptography by Sengupta et al. [132]. The cryptography is done in this work using the method of hybrid Ceaser cipher encryption. It offers protection for the cloud at the client, server, and network locations.

Somani et al.[137] suggested an RSA algorithm used by Digital Marks to boost authentication by testing it with Digital Signatures to maintain anonymity as part of security. The solution was based on encryption carried out in five steps. In the initial stage, the key was established. In the second step, advanced labelling is performed, and in stage 3 and stage 4, encryption and decoding are done. Signature confirmation is made at the last stage.

The architecture to ensure the confidentiality of information placed in the cloud by affecting the use of the computerized mark and Diffie Hellman on the Advanced Encryption Standard encryption algorithm for key exchange was addressed by Rewagad et al.[122]. Regardless of whether the transmission key is lost, in view of the fact that the travel key is of no use without the customer's private key, which is given only to the true blue occupant, Diffie Hellman's key trade office makes it pointless.

Prabhakar et al.[117] proposed an information encryption protocol with an Advanced Encryption Standard algorithm in mind. The Advanced Encryption Standard solution, using cloud mode, masks data over the entire life cycle from start to finish. To ensure cloud information exchange, this encryption approach uses an Advanced Encryption Standard-256 encryption algorithm and a Secure Socket Layer.This method prevents data from being targeted by force, and provides data across the cloud with effective protection. It doesn't concentrate much on data protection and data quality. The proposed technique ensures that data is completed in all phases and is separated into two phases. Data encryption is handled and information is transmitted safely in the cloud from the start of the stage, and the next stage handles data recovery that combines customer confirmation and decoding of information. AES-256 encryption terminates the encryption of information in the first stage. The client should be verified at the second level, and the client will send the username and the secret word to the cloud. When the cloud receives the demand of the customer at that point, it confirms the customer's subtle elements, if the customer at that point is significant, then the process of data recovery begins.

In Keiko Hashizume et.al [56], the cloud computing mechanism uses numerous advances but it also acquires its security problems which are examined in the paper, it distinguishes the fundamental susceptibilities in this type of frameworks and the most significant risks found in the writing identified with CC and its environmental condition just as it recognizes and relates susceptibilities.

Akhil Behl et.al [15] discussed nitty-gritty analysis of cloud security issues. The researchers looked at the security problem from the cloud engineering perspective, cloud-based functionality perspective, web conveyance model perspective, and cloud shareholder perspective.

R. Balasubramanian et.al [11] discussed security-related concerns in the cloud computing environment that combines the security of stockpiling, information security, and device security. The important security issue is that the data owner probably won't have control over where the data is being placed. This assumes that the benefits of using cloud mechanics must be exploited and distributed computing must be used. Scheduling and resource-based allocation provided by clouds should be used equally. In this way, there is a need to shield the information within unconfined processes of the system.

Chan YeobYeun et.al [7] proposed that only security issues for distributed computing, like IAM were important but some other factors must also be considered. In addition to establishing IAM specifications and protocols, Identity and Access Management have demonstrated the current state authorization, authentication, and evaluation/audit of clients accessing the cloud.

Grigoriev and Ponomarenko et.al [49] proposed a new definition of homomorphic cryptosystems over arbitrary finite groups such as non-Abelian groups on which homomorphic cryptosystems are constructed. It results in an exponential blow-up by repeated computations of the ciphertext lengths, as a free group product is the ciphertext space obtained from the encryption scheme. The explanation for this is that the total length of a free substance is typically the sum of the lengths x and y of the two elements x and y.

Christos Stergiou et.al [139] A study of cloud computing and IoT emphasizing the security problems of the two advances were introduced. In particular, the professionals joined the two advances (e.g. IoT and Cloud Computing) to examine the unchanging highlights and figure out the mixed rewards. In this way, it helps to determine how the cloud computing system improves the IoT dimensions.

Preeti Mishra et.al [101] gave a comprehensive overview of the various intrusion detection strategies proposed for cloud environments with an investigation of their capability to recognize attacks. In cloud dynamics, the researchers have suggested a risk model and taxonomy of attack classification to explain the susceptibilities or vulnerabilities.

Pallavi Meharia et.al [99] proposed methods for promoting a secure communication mechanism between individuals and their extensive devices through the application of data based on the human physiological system as a recognition criterion. Various biometric strategies were examined, and the basis behind the applicability of the system contended.

Prathamesh Churi et. al [29] proposed a new algorithm for improvising password encryption using Jumbling-Salting-Hashing. One of the most serious password security problems is securing encrypted passwords in the server database. The most popular methods of guessing passwords in cryptanalysis include a dictionary attack or brute force attack. The jumbling process consists of randomly selecting and adding characters from the predefined character set to the plain password; salting consists of prepending a random string and hashing is performed to obtain a fingerprint stored in the server database using the cryptographic hash feature.

Alejandro Velez et. al [150] proposed a new technique to strengthen the ciphertexts produced in double random step encrypting experimental configurations. This ciphertext was secured by multiplexing with the same setup coded with a "salt" ciphertext. An experimental implementation of the technique of "salting" was presented in this paper. In some of the widely recognized attacks the resistance of the "salted" ciphertext, demonstrating the validity was examined.

2.3.1 Data Encryption Mechanisms

Many of the existing architectures have attempted to use various cryptographic encryption techniques to provide their data owners with a stable, secure cloud environment to store their most sensitive or highly confidential data. The works in [42, 91] proposed a novel cryptographic technique that is a reflection of the Ceaser cipher where substitutions with eight-bit manipulations are used. Until being encrypted, the process eliminates the repeated words in the plaintext to make it unreadable for an attacker to analyze the original plain text from the encrypted one. For transformation operations, Amit and Bhavesh suggested the randomized cryptographic method [158]. Using both randomized and public key generation methods, they corrected different variants of Ceaser cipher substitution techniques and a transformation protocol for its implementation. On different plain text sizes, they checked their algorithm and found that the system worked well on broad plain texts. Quist and AphetsiKester presented a hybrid approach [110] for the encryption as well for decryption operations, where they combined both classical substitutions and transposition techniques. An initial columnar transposition cipher is applied on the plain text to encrypt it, and next, on the resulting cipher, the method applied the Vigenère cipher technique to finish the final encryption of the data. The method used a column transposition, where a fixed column length key is chosen, and the plain text is arranged row by row in the chosen table of fixed column length. Finally, the approach rearranged the columns in alphabetical order of key and read the resulting text column by column. This final output cipher is taken as a key for the vigenere technique.The mechanism for the decryption process is implemented in the same way but in the reverse order. The presented work illustrates the performance of vigenere cipher that is known for stronger ciphers that uses the strength of columnar transposition.

Fadhil highlighted a hybrid technique in [147] by combining the features of the public-key cryptosystem RSA and that of the process of greedy knapsack to provide a highly protected and less complex device. The plain text is encrypted using the RSA algorithm in the first phase, and the resulting output cipher is used as input for the greedy knapsack algorithm in the second phase. The decryption order is reversed at the receiver. The performance of the method is observed in a way the hybrid method took less time in comparison to RSA and knapsack for encryption and decryption of the plaintext. Works discussed in [69] proposed a hybrid combination of substitution and combination techniques to remove their weakness and finally they developed an encryption scheme strong enough to produce a strong cipher. The work used the combined techniques of Ceaser cipher substitution and Rail fence transposition to increase their efficiency. Works presented in [135] showed a fully homomorphic encryption technique developed by IBM in June 2009. The approach showed the processing of encrypted data without being decrypted. Roy and Ramadan used decentralized information flow control (DIFC) and differential privacy security techniques in [125] in the mechanism of data generation and different cloud calculation stages and came across a privacy protection system called AIRAVAT. Studies showed the method performed well for privacy leakage prevention.

2.3.2 Key Management Methods

A key problem for cloud data security is key management, how various keys that are generated can be efficiently managed. The problem has been raised as the users are not so experts to manage their keys. Now it is the burden of the cloud service providers to maintain an efficient mechanism to maintain these large number of user keys. The efficiency is most concentrated here because there are multiple numbers of users and multiple numbers of prioritized classes. To increase its efficiency, the cloud service provider has developed its Key Management System (KMS). If the system has multiple users and prioritized classes, the Key Management System (KMS) after generation of the keys has to distribute the keys to the users. If the key is compromised, KMS has to recover the key. KMS has to identify unused keys and delete them.

KMS has to perform the integrity check of these keys.

Many of the cloud Key management techniques are associated with the metadata. Metadata maintains information about Key labels, Key Identifiers, Key lifetime, cryptographic parameters that generated the key, key length, and key usage count. Works discussed in [76] presented Key life cycle concepts, describing various stages such as key distribution, key active state, inactive state, and termination. The work presented a general Key management taxonomy wherein key revocation and verifiability are focused. A key management approach at the client side is discussed in [2, 76]. The approach used homomorphic images to manage the keys on the client side. If the key is lost, the user can manage a new key with the metadata at his side. The homomorphic approach proved efficient in managing and identifying unique user keys.

Works discussed in [62] showed key management at the Cloud service provider's side. Here if the key is lost, the user cannot see his stored data in the cloud. The user has to again request a new key from the service provider. The service provider generates a variable new key to the customer at his request. Works in [50] showed a useful key management approach at both the user and service provider side. In this approach, the key is divided into two parts where on the user side the first part is stored, and the other part is stored on the cloud service provider side. User data can be retrieved by combining the two key parts. If any part of the key is compromised data cannot be retrieved, and a new key has to be generated and maintained. Works showed this is an efficient management scheme as the data owner can trust that even the cloud provider cannot access his sensitive data.

Key Management using a Centralized Server is discussed in [3, 66, 86]. The approach used a public key cryptosystem wherein a pair of the private part of the key and the public part is managed. The key management centralized structure as the cloud uses the public part to encrypt the user data. This encrypted data can only be decrypted by the user's private key which is managed by the user himself.

Works discussed in [23, 103, 119] presented a Group Key Management mechanism. The service provider identifies a few trusted members of the cloud and forms a group. The group key is formed using the private keys of each of the group members. Using the group key data can be accessed. If any of the group members leave the group, then a new group key has to be generated and managed. In [28] the author discussed how the advent of the Organization for the Advancement of Structured Information Standards (OASIS) with the help of the Key Management Interoperability Protocol (KMIP) is successful in solving the key management issues. A better approach used for efficient key management is class data integrity verification. An integrity check is an efficient key management method where the keys are managed based on the integrity check codes generated on the data works discussed showed that NEC Lab's validated data integrity (PDI) solution could efficiently help verification of data integrity at finer granules. Cong Wang suggested a mathematical approach to dynamically check the integrity of the data stored in the cloud in [152]. The model generated mathematical checksum codes to identify the keys used for data encryption.

2.4 Comparative analysis of existing techniques

The analysis of existing techniques has been demonstrated in the below Table 2.1 and Table 2.2:

Table 2.1: Comparison of various Classification Techniques

Author's Name	**Technique**	**Finding**	**Limitation**
[113]	Decision Tree	It can deal with noisy or incomplete data quickly. It can handle continuous and discrete data, both.	The classification error rate is high, while the training set is limited relative to the number of classes.
[45]	Neural Network	This algorithm can work with incomplete knowledge and having high fault tolerance.	There is no specific rule for determining the structure of artificial neural networks. It is based on experience and trial and error.
[97]	Naive Bayes	It is easy to implement and always provides good results.	Dependencies among variables cannot be modeled.
[130]	SVM	If there is no idea of the data, SVM's are very good and it works well for semi-structured data such as text, images, and trees.	It takes a long training time for large datasets. It is difficult to understand and interpret the model.
[159]	K-NN	It is robust to noisy training data. It is very if the training data is large.	Distance-based learning is not clear and the computation cost is high.

Table 2.2: Comparison of various Encryption Techniques

Author's Name	**Technique**	**Finding**	**Limitation**
[137]	RSA	This algorithm is very safe and secure for the users and it is very fast and simple.	It is not a scalable technique. Security is applied on user side only.
[146]	Reverse Ceaser	It is a scalable approach and security is applied to user data.	It is considered a week method of cryptocurrency, as it is easy to decode the message.
[103]	RSA and Diffie Hellman	It provides effective data security with low overhead and communication.	The major drawback of this work is user accountability.
[122]	AES	This approach prevents data from brute force attacks and provides effective security to data.	It does not emphasize more on privacy and efficiency of data.
[25]	Blow Fish	This algorithm is one of the fastest block ciphers in general use.	In Blowfish algorithm, key management is very complicated and does not provide authentication when two people have the same key.

2.5 Research Gaps

This section contains various gaps found in the existing literature.

i. Most of the existing encryption techniques emphasize encryption time only and the storage space is ignored.

ii. It has been observed that the size of the encryption data has not been optimized.

iii. It has been observed from the literature that the unclassified data is not monitored so the proposed work will focus on the unclassified data.

iv. The majority of existing classification techniques have ignored accuracy, precision, and recall in the analysis.

2.6 Problem Formulation

More knowledge that has been produced in the past two years than in the entire previous history of the human race is overflowing in data volumes. Photos, videos that transfer over social media are not safe from eavesdroppers. Protection of big data is required on social media. But here providing security of the confidential data is the key issue and it could be handled with encryption. But encryption is usually combined with padding, which will increase the size of the data. Therefore, the main necessity is to design an integrated approach to classify and encrypt the data sending it to the cloud.

To overcome the different problems with existing techniques, a novel classification technique will be proposed in this research work. The proposed plan will handle the big data with classification and encryption. Firstly data will be classified into a sensitive and non-sensitive category and the encryption of only sensitive data will be done. The encryption time and the storage will be challenging part because storage always increases after encryption. So the size of the encryption data should be optimized.

2.7 Objectives

i. To design and develop a classification technique for reducing the error in the prediction of the sensitive data.

ii. To enhance the existing encryption approach by minimizing the time and storage space.

iii. To compare the proposed framework on the basis of classification and encryption techniques.

The parameters that will be used to analyze the performance of the proposed system:

Precision, Recall, and Accuracy in classification Technique.

Computation time and storage space of encryption Technique.

CHAPTER 3

SENTIMENT ANALYSIS AND HOMOMORPHISM FUNCTION PROPERTIES

Outline

In the era of information technology, boundless data is available on the Internet. Public and private organizations can take better business decisions and succeed tremendously if they use the information available on the internet. For instance, they can get to know how customers are liking their product. But the information or data available on the internet is not organized; it is in raw form. To make sense of this Big Data, we need to apply preprocessing techniques to identify several characteristics and do Sentiment Analysis.

3.1 Need of Sentiment Analysis

Sentiment Analysis is an art of raw data analysis to generate information that is both useful and beneficial in several respects. Data Analytics is now widely used for better business decisions in many organizations. Through applying analytics to structured and unstructured data, the companies introduce a lot of improvements in the way they prepare and make decisions. If we can make good use of those data, we can take a lot of advantages from that. About 80 percent of digital data in the world is unstructured, and data from social media outlets is not useful. Since the information is not structured in any predefined way, anything from it is difficult to sort and analyze. Models can now be built that learn from examples and can be used to process and organize text data, thanks to the advances in Machine Learning and NLP.

Big Data analytics systems allow us to categorize large sets of data into sensitive and non-sensitive data and detect the polarity of each argument automatically. And the best part is that it's fast and simple, saving valuable hours and allowing us to concentrate on projects where we can make a greater impact.

The parameters for Sentiment Analysis refer to the use of NLP, text analysis, and computational linguistics for information classification and extraction. Sentiment Analysis requires analyzing the text's feelings or behaviours of the writer. To accomplish this mission, Opinion mining utilizes data mining and machine learning concepts.

Sentiment analysis is commonly used for several uses in social media feedback. Analysis of the sentiment aims at determining the overall polarity of the results. Daily, large amounts of data are generated from social media and distributed to the World Wide Web.

Most companies are using opinion mining systems to test the opinions of various customers on sold goods. Opinion mining is a new and different way to keep multiple business trends in focus. This data contains very critical information that can be used for business and other scientific industries to benefit. Manual extraction of this valuable information is not feasible due to large data that is being generated in massive amounts daily, so it is important to evaluate sentiment. Sentiment Analysis is the practice of extracting useful knowledge, feelings, or opinions from user feedback on a given topic, area, or product. It extracts information from the source data. It is an application of NLP, computational linguistics, and text analysis. It clusters the feelings into such categories as positive or negative. Analysis of social media feelings can be an excellent source of information. Companies are gradually using the content of these social media sites for different purposes, with the accelerated growth of Big Data on the Internet for social media.

Today, if anyone wants to buy some product, that person is no longer limited to asking for opinions from his friends and family. As there are many user reviews and conversations about the project on the Internet. This can no longer be appropriate for an organization to perform surveys, opinion polls, and focus groups to collect public opinion since such knowledge is already available.

3.1.1 Sentiment Analysis Classification Approaches

The sentiment analysis can be performed using the three main approaches. They are as follows:

- **Lexicon Based Approach:** It is based on a lexicon that is often referred to as a dictionary approach and forecasts a pre-calculated polarity lexicon or dictionary terms. Using a manual dictionary or any other available corpus, the lexicon can be produced.

- **Machine Learning Approach**: It requires a corpus with a wide range of tagged examples. Using these examples the machine learning tasks involve learning methods.

- **Hybrid Approach**: The combination of machine learning and lexicon methods is known as a hybrid approach.

- **Deep Learning Approach**: Deep learning algorithms run data through several "layers" of algorithms in the neural network, each transferring a simplified representation of the data to the next layer.

3.1.2 Steps to Classify Text

There are two categories of classifiers: they have supervised classification and unsupervised classification. All the classifiers like trees, rules, lazy, and naïve come under these categories [12]. This is best suited for classification of the given data and visualizing the output. And the classification is not done in a single step. A sequence and several steps are involved in classification.

- Data preprocessing, preparing the data for classification
- Create training dataset
- Choose Classifier
- Train the training dataset
- Analysis of the result or output for performance

Basic functionalities in text classification

The basic functionalities required in text classification process is given below:

- **Data preprocessing:** Raw data from social media is highly noisy. These noisy data may contain URLs, sarcasm, emoticons, other language alphabets, abbreviations, and so much more. This noisy data is often incomplete and inconsistent. Analyzing these data mostly leads to errors. Any text for classification should be preprocessed or cleaned before using it. It is similar to the tasks involved in a database like data cleaning, data integration, data transformation, and many others. Depending on the domain, the noisy raw data is preprocessed.

- **Classification:** Classifiers are separated into rule-based methods, tree learners, function-based learners, and incidental techniques.

- **Clustering:** It is a data mining technique that is used to group abstract or physical objects into different categories of similar types. It is unsupervised learning that can group instances of given unlabeled data. It is a subset of similar objects. It is the task of subjectivity. There are various algorithms like k-means, hierarchical, and mean shifting clustering.

- **Attribute selection:** The dataset often includes a huge number of attributes or features. The technique of extracting only the relevant subset of attributes is called feature selection. It could be done separately or combined with the learning process. This can enhance the interoperability of the model selected, speed the learning process, and improve learner performance. There are various techniques available to identify or select the features.

- **Data visualization:** Information can be assessed outwardly by plotting attribute values against the class. Classifier yields can be contrasted with preparing information keeping in mind the end goal to recognize anomalies. For particular strategies, there are specific representations. Weka tool also includes support for association rule mining, comparing classifiers, data set generation, facilities for annotated documentation generation for source code, distribution estimation, and data conversion.

3.1.3 Elements used to Generate Result

The various elements that are used to generate the result are listed below:

Mean Absolute Error (MAE)

In a series of predictions, the MAE calculates the average magnitude of the errors, without taking their path into account. For permanent variables, it tests accuracy. MAE is the average of the absolute values of the contrasts between prediction and interpretation relative to the confirmation test. Mean absolute error is given by the equation

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

$$x_i = actual\ value \quad n = sample\ size$$

$$y_i = predictions$$

Root Mean Squared Error (RMSE)

The RMSE (Root Mean Squared Error) is a rule of quadratic scoring used to calculate the mean magnitude of the error. The RMSE gives a comparatively high weight to the considerable errors.

$$\text{RMSE} = \sqrt{1/\text{N} \sum_{i=1}^{N} (\hat{\theta} i - \theta i)^2}$$

The Root Mean Square deviation of the predicted values $\widehat{\theta i}$ is observed values and θi is modelled values at a regression dependent variable time/place i.

Relative Absolute Error (RAE)

Basically, the relative absolute error is the same as the relative square error, since it is applied in relation to the simple indicator, which is the average of the real values. Numerically, the relative absolute error i of an individual program 'i' is assessed by the condition:

Where P(ij) is the esteem predicted by the individual program i for test case j (out of n test cases); T$_j$ is the objective incentive for test case j; and is given by the equation For a perfect fit, the numerator is equivalent to 0 and E$_i$ = 0. In this way, the E$_i$ record ranges from 0 to infinity, with 0 relating to the perfect.

$$\mathrm{RAE} = \frac{\sum Ni = i|\hat{\theta}i - \theta i|}{\sum Ni = 1|\bar{\theta} - \theta i|}$$

Root Relative Squared Error (RRSE)

The root's relative square error is relative to the simple predictor which was used. By taking the square root of the relative square error, one reduces the error to an indistinguishable calculation from the estimated amount.

Mathematically, the root relative squared error Ei of an individual program i is assessed by the equation

$$\mathrm{E}_{ij}\frac{\sqrt{\sum_{j=1}^{n}\left(P_{(ij)} - T_j\right)^2}}{\sqrt{\sum_{j=1}^{n}\left(T_j - \bar{T}\right)^2}}$$

Where P(ij) is the value predicted by the individual program i for test case j (out of n test cases); Tj is the target value for test case j;

and is given by the equation below

$$\bar{T} = -1T_j$$

The numerator is equal to 0 for a perfect fit and Ei= 0. The Ei list ranges from 0 to infinity in this way, with 0 compared to the ideal one.

Kappa Statistics

Cohen's kappa coefficient is a statistical measure that calculates inter-rater agreement for qualitative things. Kappa measures the agreement between two raters, each arranges N things into C fundamentally unrelated classifications. The principle behind this kappa-like statistical parameter is ascribed to Galton (1892)[80].

The condition for k is K $= \frac{P0-Pe}{1-Pe}$

Where 'P 0' is the relative observed agreement among raters, and 'P e' is the theoretical likelihood of chance agreement, utilizing the observed information to compute the probabilities of every observer randomly for every classification. On the off chance that the raters are in total understanding then k = 1. Sometimes there is no understanding among the raters other than the normal by possibility, where K ≤ 0.

3.2 Homomorphic Encryption Scheme

A Homomorphic encryption scheme is considered to be one of the trending methods used to provide security of data in the cloud. This procedure is applied to the cloud provider. Initially, the original message is encoded by using the encryption procedure. The cipher text that is generated is stored in the cloud server. At this juncture; there is a possibility of this cipher getting hacked by the intruders. This has to be handled. Hence, Homomorphic encryption plays a prominent role in providing security for data in the cloud. The process of the Homomorphism encryption scheme is that it performs encryption on the encoded information stored in the cloud. Homomorphic operations are performed on the enciphered information. The result obtained from the encryption procedure is stored in the cloud with the goal that nobody else has an access to the information that is stored in the cloud server. Consequently, security is improved. Homomorphic encryptions can be either partially homomorphic or fully homomorphic. If the scheme employs either multiplicative or additive operations on the encrypted data then it is a partially homomorphic encryption scheme. If both of the multiplicative and additive operations are applied to the enciphered data then the scheme is fully homomorphic. RSA algorithm and Elgamal follow asymmetric cryptography. Both the algorithms satisfy only multiplicative operations. Hence they are partially homomorphic [72].

3.2.1 Homomorphic Property

The homomorphic property is defined as: G is a divisor of “x1” and “y1”. Then, GCD(x1 ,y1) = GCD (x1, -y1) = GCD(-x1, y1) = GCD(-x1,-y1).

Generally, GCD(x1, y1) is as follows:

GCD(x1,y1) = GCD(mod(x1), mod (y1)).

$$A = q * n + r \ , \ 0 < = r < n \ , \text{ and } q= [x/ n] \ ,$$

where (x/n) is the greatest number which is an integer and should be less than or equal to [x / n] , “q” is the quotient with “r” as remainder .

3.2.2 Boolean function

Let us consider $[a_1, a_2 \ldots\ldots a_n\}$ to be the set of attributes such that $\{(a_1 \vee a_2 \vee a_3) \wedge (a_4 \vee a_5) \wedge (a_6 \vee a_7)\}, y : \{0, 1\}^n \to \{0, 1\}$. The adaption scheme for the given binary number finite set will be modified into single monotone function 0, 1 for a given monotone M program over a field F of r × c matrix where r denotes rows and c denotes columns with a renaming function q: $[s_1] \to [s_2]$; then, the adaption scheme: A_p is s $(e_1, e_2, \ldots e_n)$ E 0, 1, thus $A_p(e_1, e_2, \ldots e_n) = 1$, where e1, e2 . . . en are set of elements.

3.2.3 Inversion function

The A_p adjustment policy for a finite set of reverse binary numbers is changed to a single monotonous function 1, 0 for a given reverse monotonous programme M^C over a matrix field F or r x c with an inverse function q^{C} : $[\mathrm{s}_1] \to [\mathrm{s}_1^{\mathrm{C}}]$; and then, the strategy of inverted adaptation:

$$A_p^c \text{ is } (e_1^c, e_2^c \ldots e_n^c)\, E\{0, 1\}^n$$

3.2.4 Left shift (Ls) & right shift (Rs) function

Policy for Adaptation For a given LS or RS monotone function M « K or M », Ap of the finite set of right or left shifted binary numbers will be changed into a single monotone function 1, 0 N, where N is the value of a given constant over a field F. An r × c matrix with a left shift LS function q1: [s1] « [s2] and right shift RS q2: [s1] » [s2] where the adaption policy L_S or R_S is $(e_1, e_2 \ldots e_n)E0, 1^n$.

3.3 Summary

The main sentiment analysis approaches are the lexicon-based approach, Machine Learning approach, hybrid approach, and Deep Learning approach. General steps for classification of text are (i) data preprocessing, (ii) Creating training and test datasets, (iii) choosing the classifier, (iv) training the training dataset, and (v) analyzing the result for performance measurement. Various model evaluation parameters are mean absolute error, root mean squared error, relative absolute error, root relative squared error, kappa statistics. One of the trending methods used to provide security of data in the cloud is the homomorphic encryption system. This procedure is applied to the cloud provider. It performs encryption on the encoded information stored in the cloud. Homomorphic encryptions can either be partially homomorphic or fully homomorphic.

CHAPTER 4

SENSITIVE ENCRYPTED STORAGE

Outline

Sensitive Encrypted Storage (SES) is a proposed framework to improve the encryption of Big Data by intelligent classification techniques. In this Chapter, an optimal solution and methodology is proposed to classify the social networking text into sensitive and non-sensitive segments. Moreover accuracy of various classification algorithms is compared and the best one is chosen for the further process. Further, only the sensitive data is encrypted using proposed ECDH-SAHE Encryption method to save encryption time, space and storage.

4.1 Proposed Methodology

Many algorithms use different parameters to prove efficiency. The parameters could be time, space, cost, and accuracy. Using an intelligent classification strategy, the proposed approach segments the big data with precision. An alternative approach is designed to decide whether it is appropriate or not to break data packets to shorten running time and minimize storage space.Our experimental evaluations of both safety and efficiency performance and experimental results show that our method can resolve major cloud hazards effectively and that using an intelligent machine learning classification technique requires an appropriate computation time. We suggested a new solution entitled Sensitive Encrypted Storage(SES) as a model.

In Figure 4.1, the flowchart of the proposed model is shown. We used our proposed algorithms in this model, including Logistic Regression Convolution Neural Network (CNN-LR), Elliptic-curve Diffie-Hellman-Shifted Adaption Homomorphism Encryption (ECDH-SAHE), and Elliptic-curve Diffie–Hellman-Shifted Adaptation Homomorphism Decryption (ECDH-SAHD).

4.1.1 Phases of Proposed Methodology

The methodology is classified into three phases and each one is described below:

Phase 1: This stage of the proposed work will be focused on the model of secure data classification, which will further be based on the data level of sensitivity and will be categorized according to that level. Precision, recall, and accuracy will be evaluated in this phase.

Phase 2: This stage will encrypt and store sensitive data in the cloud and the same cloud is used to store non-sensitive data for the efficient use of data. In this phase, computation-time and storage space of the encryption technique will be analysed.

Phase 3: The aim of this phase of the proposed work would be to provide better results than the current algorithms through the use of precision, time, and confidentiality and integrity of cloud data parameters, and ultimately to minimize overall execution time and total costs.

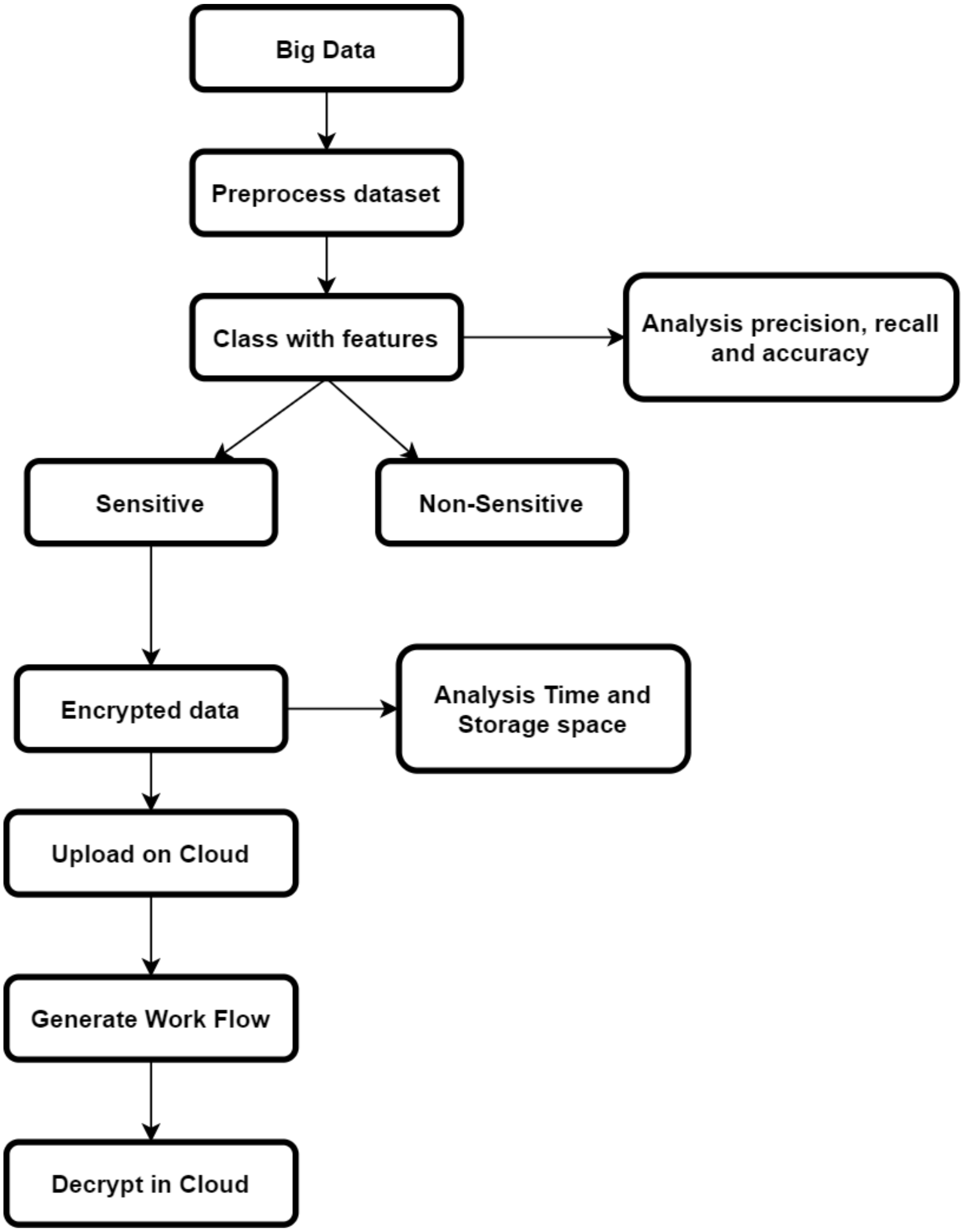


Figure 4.1: Graphical representation of the proposed technique

4.1.2 Proposed Framework

The proposed framework is shown in Figure 4.2 and Figure 4.3. The process of data classification and Encryption is illustrated with the help of figure 4.1. Big data will be classified into sensitive and normal data using novel classification approach. The next phase will work on the novel Encryption approach which will encrypt the classified sensitive data only. The sensitive data will be store on Virtual Machine-1 (VM1) and the normal data will be store on Virtual Machine-2 (VM2).
In Figure 4.3, data retrieval, decryption and data merging process is shown. The encrypted data stored on VM-1 will be decrypted using novel decryption approach. After that the decrypted sensitive data and the normal data will be merged.

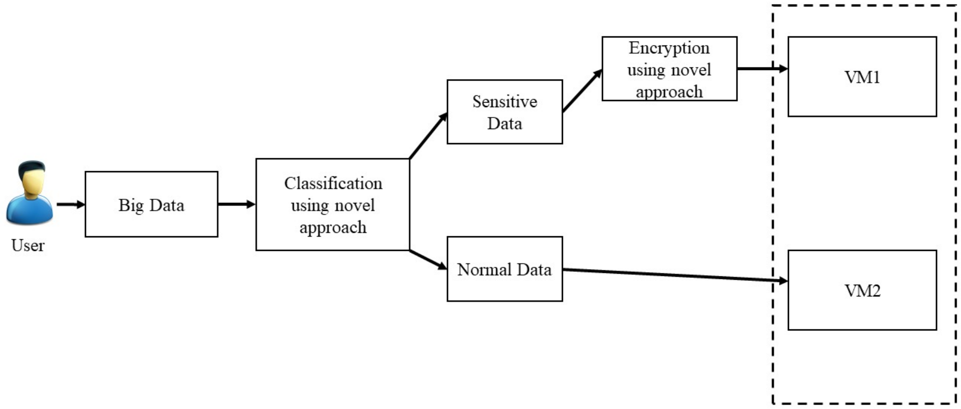

User
Big Data
Classification using novel approach
Sensitive Data
Normal Data
Encryption using novel approach
VM1
VM2

Figure 4.2: Classification and Encryption process

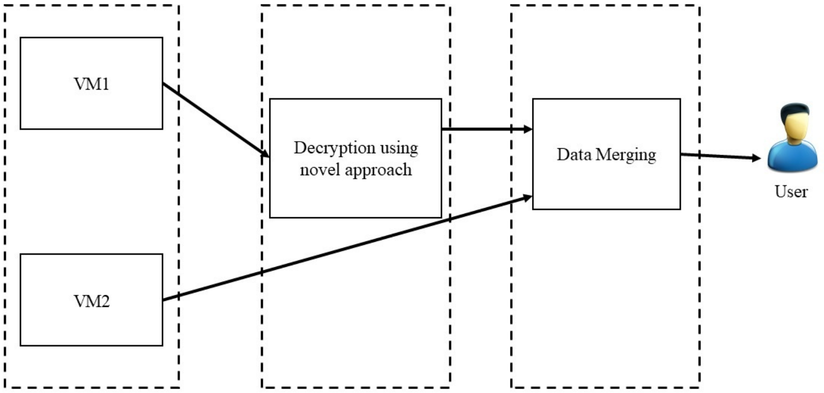


Figure 4.3: Decryption process

4.2 Sentiment Analysis Using CNN-LR Algorithm

This classification of text is a statistical model. This is based on the probability of the occurrence of a word in a sentence. Each word in the sentence is given a weightage, several times occurring in the sentence. This is the seed word or called the trained dataset. The probability of occurrence of the test data is calculated based on the trained data. The sentence with high frequency is classified for its polarity. The model is simple and very easy to develop. If each individual word is taken for calculation, then it is a unigram approach. Similarly, two words could be taken into consideration called bigram. If three words are taken, then trigram and if n-words are taken, then n-gram approach. This research starts with a unigram approach. Given a document D, and the equation is given by D (c|y) from P (c), D (y) and D (y|c). The equation is given as

$$D\left(\frac{c}{y}\right) = \frac{D(y \mid C)D(c)}{D(y)} \tag{4.1}$$

Where,

- D(c|y) is the probability of class (c, target) given predicator y.
- D(c) is the prior probability of class.
- D(y|c) is the possibility, which remains the probability of predictor given class.
- D(y) is the former probability of predictor.

4.2.1 Data Preprocessing

A tweet's extracted text is raw data containing several needless characters, symbols, and stop words that must be eliminated using natural language processing techniques. Several NLTK functions have been used for the preprocessing of the data. During preprocessing point, key information was first extracted from tweets and all the unnecessary content was removed. The NLP pre-processing methods used for the method being proposed are:

Noise removal

Noise reduction is done very carefully as it often removes a few numbers from the dataset line, leading to decreased accuracy. In data set cleaning, the basic expression used was capable of eliminating unused white spaces.

Removing Stop Words

The Stop words are the most common words in English that have little relation to the study of feelings. Some of the stopwords are "are", "the", "at", "of" etc. and these must be removed.

Repeated letter removal

Convert two or more letter repetitions to two letters.

To highlight some words, some people send tweets such as **I am sooooooooo happppppy** by adding multiple characters to a single word. This tweet,for instance, is converted to **I'm soo happy** to handle them.

We do some general pre-processing on tweets first, which is:-

- Turn the tweet to lower case.
- Replace space with 2 or more dots.
- From the ends of the tweet strip the spaces and quotes (" and ').
- Replace 2 or more spaces with a single space.

URL

Users also post tweets containing hyperlinks to other websites. For text classification, any particular URL is not relevant, as it would result in very sparse functions. So we replace all of the URLs in tweets with the word URL. The standard expression used for corresponding URLs is **((www\.[\S]+)|(https?://[\S]+))**.

User Mention

Every Twitter user has an associated name. Users sometimes reference other users by @handle in their tweets. We substitute the term

USER MENTION for all the user mentions. The standard expression used to suit the name of users is @[\S]+.

Emoticon

Within their tweets users often use a variety of emoticons to express various emotions. It is difficult to balance all the different emoticons used on social media exhaustively, because the number is enormous. Nonetheless, we match some popular emoticons that are used quite often. We substitute the corresponding emoticons with either EMO POS or EMO NEG, depending on whether they express a positive or a negative emotion.

Hashtag

Hashtags are non-spaced phrases prefixed with the hash symbol (#) that users often use to identify a trending subject on Twitter. We remove the # symbol for all hashtag words. For example, replacing # hello with hello. The standard hashtag matching expression is # (\S+).

Retweet

Retweets are tweets that someone else has already posted, and that are shared by other people. Retweets start at the letters RT. We delete RT from the tweets as this is not an important function for classifying messages. The standard expression used for corresponding retweets is \brt\b.

After applying tweet level pre-processing, we processed individual words of tweets as follows:-

- Strip any punctuation [’"?!,.():;] from the word.
- Remove - and ’. This is done by converting them to the more

general words like shape of the shirt and theirs to handle words such as t-shirt and theirs.

- Check to see if the word is valid and only accept it if it is. We describe a valid word as a word that starts with alphabets, numbers or a dot(.) and underscore(_) being successive characters.

Table 4.1 shows some of the unnecessary contents which were removed from the main tweets.

Table 4.1: Some removed contents from original tweet

Contents	**Actions**
Punctuation	Removed
Uppercase Character	Lowercase all contents
All word	Converted into simple form
Empty space	Removed
Number	Removed

Therefore, to create a dataset that can be easily learned by different classifiers, raw Twitter data has to be normalised. To standardise the dataset and decrease its size, we have applied an extensive number of pre-processing steps.

4.2.2 Feature Extraction

The extraction of features is an important step when dealing with natural languages because a computer does not understand the text we have collected. When we have a tweet, anything like this goes:-

I do not like the views of @some1 on #Topic1. Too conservative!! I can't stand it!

We can't feed this into a learning algorithm. We need to convert it to a proper format, so we'll pre-process our data.

We may want to start by trying to tokenize the message. Tokeniszation is like segmentation. In Python, libraries like NLTK are already available for such functions. Transformation can be done according to how tokens are made:

[‘I’, ‘do’, ‘not’, ‘like’, ‘the’, ‘views’, ‘of’, ‘@’, ‘some1’, ‘on’, ‘#’, ‘Topic1’, ‘.’, ‘Too’, ‘conservative’, ‘!’, ‘!’, ‘I’, ‘can’, ‘“”’, ‘t’, ‘stand’, ‘it’, ‘!’]

Tokenization

This move separates the large paragraphs into tokens that are phrases. Such phrases may also be divided into sentences. Consider an example, before tokenization “The doctorate is a tough work to do" and after tokenization it becomes {‘?’ ‘PhD’,“tough,”work,“to”, ‘do’}

Normalization

To achieve normalization there are several activities performed simultaneously. This involves translating all text into upper or lower case, removing punctuations, and converting numbers into their corresponding words.

Stemming

This method is also helpful in eliminating unnecessary word computation by using the stemming technique to transform different tenses of words into their base form. For example, fishing, fishing, fishing to fish, arguing, arguing, arguing

There is much more to extraction of the features. Here are some things to keep in mind:

- Instead of counting terms, we can go with **frequency** (i.e. number of occurrence of a term divided by total number of terms)

• Consider reducing the number of unique words or applying weighted schemes (like **TFIDF**)

Text files are simply words (ordered) in sequence.In order to run machine learning algorithms, we need to transform text files into numerical feature vectors. We're going to use the bag of words model for our example. In short, we divide each text file into words (for space-splitting English) and count number of times each word appears in each document, and then assign each word an integer ID. In our dictionary, each specific word corresponds to a feature.

Scikit-learn has a high-level element that will generate 'feature vector'.

TF: TF (Term Frequencies) is frequency term i.e. #count(word) / #Total words, in each document.

TF-IDF:Finally, the weighting of more common words such as (the, is, and) that appear in all documents may also be reduced.

N-Gram

The text characteristics for supervised machine learning algorithms will be shaped by N-Gram. This is a sequence of n tokens in the given text. The value of n could be 1, 2, 3, etc. For n = 1, it is called unigram, for n=2, bigram, for n=3 trigram, and so on.

If we take a sentence into consideration.

"Lovely is better Institute". If we consider N = 2 then it will produce "Lovely is", "is better", "better Institute".

Two types of functionality, namely unigrams and bigrams, will be extracted from the dataset. We create a frequency distribution of the unigrams and bigrams in the dataset that are present and select top N unigrams and bigrams for analysis.

Unigrams

The inclusion of single words or tokens in the document is possibly the easiest and most widely used feature for classifying text. In the training dataset, we extract single words and establish a frequency distribution of these words.

Bigrams

These are word pairs that occur successively in the corpus. These features are a good way to model negation in natural language, as in the word –, which is not good. We retrieved a total of 1954953 separate bigrams from the dataset a total of 1954953 separate bigrams. Among these, the noise at the end of the frequency spectrum is one of the bigrams and occurs very few times to impact classification.

4.2.3 Training Dataset

Supervised learning is an important technique for solving problems concerning classification. Training the classifier allows the use of unknown data for possible predictions. Here we will train dataset of one lakh tweets.

4.2.4 Applying Classifiers

For this research work, different types of machine Learning classifiers for performance evaluation are applied.

Cross Validation

This technique is used to measure the predictive performance of a selected model. The sample dataset is partitioned into a training set to train the model and test dataset for evaluation. It is used to validate the chosen model for prediction. It is also called rotation estimation as the validation is performed in different folds. It is mainly used in the problem of prediction, to estimate how accurately a predictive model will perform in practice. There are different types of cross-validation methods. Here in this research three types of validations are used.

K-Fold cross validation

In this validation, the original dataset is partitioned into k equal size datasets. Randomly one dataset is retained for validation or called test data and the remaining k-1 sets are considered as a training dataset. This process of training the data and validating the data is continued by repeatedly k times until all sets have been considered as a training dataset and test dataset. The error rate in each iteration is then averaged to get the final error.

Holdout method

It is one of the simplest validation techniques. The original dataset is partitioned into two sets. One set is taken as the test dataset. The other set is the training dataset. Then a function approximator is used to predict the value of the testing set. The errors accumulated are precisely calculated in this validation technique.

Leave One Out Cross Validation (LOOCV)

LOOCV is a validation technique similar to the k-fold cross-validation method. The difference is partitioning of the dataset is equal in size in the k-fold technique, whereas the size may not be the same in the LOOCV method. One set is considered as the test dataset and the remaining dataset is used for the training set. The disadvantage of this technique is the expensive cost in implementing, as it has to be repeated a large number of times.

4.2.5 Implementation of CNN-LR Algorithm

The implementation of the system will start by choosing a validation method, k-fold cross-validation technique. Here a ten fold cross-validation is done. Once the validation is finished, the probability of the word occurrence is calculated. The flowchart of an implementation of the algorithm is given in Figure 4.2 and 4.3. This system is a 2-class classifier that classifies the given tweets for their polarity as positive or negative. The tweets which the classifier was not able to identify are termed as neutral polarity. This is also called a unigram approach since each word in the given tweets is taken for the calculation. The trained dataset consists of words and weightage associated with them [46].

In the first step, the input text is separated by full stop and processed as a sentence. In step second, pre-processing of the text in the form of computer understanding is performed and reduction of noise from the text is done. In the first step it changes the sentences in words, and in the second step, it removes the full-stop, comma, and exclamatory signs and in the third step it uses stemming approach which reduces the noise from words. In this approach it changes the words to their original form, for instance, if any word ends in "ing", it will be changed into original words. The next step of architecture changes the words vector into the space vector which is found by the frequency of words. After that, makes a feature matrix and labels the classes and learns to classify and analyse the parameters. This architecture analyses the classification models using machine learning or deep learning approaches.

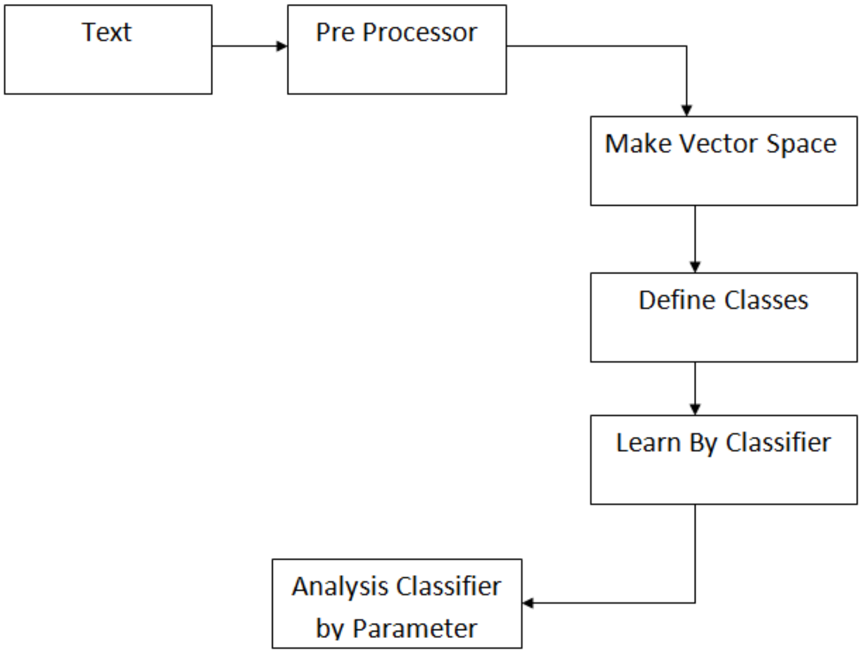


Figure 4.4: Architecture of Text Classification.

Algorithm 4.1 :Pre-Processing

Input: Text

Output: Vector words with class label [+1,-1]

```
While (Number of Text > 0)
Start
Tokenize the text in words
Remove stop words
Apply Stemming
End
```

Algorithm 4.1 gives brief description of pseudo code of pre-processing steps of text which explains the architecture part of the work. In this pseudo code, all the text is tokenized word by word and then stop word removal is performed for noise reduction. Then all words are changed

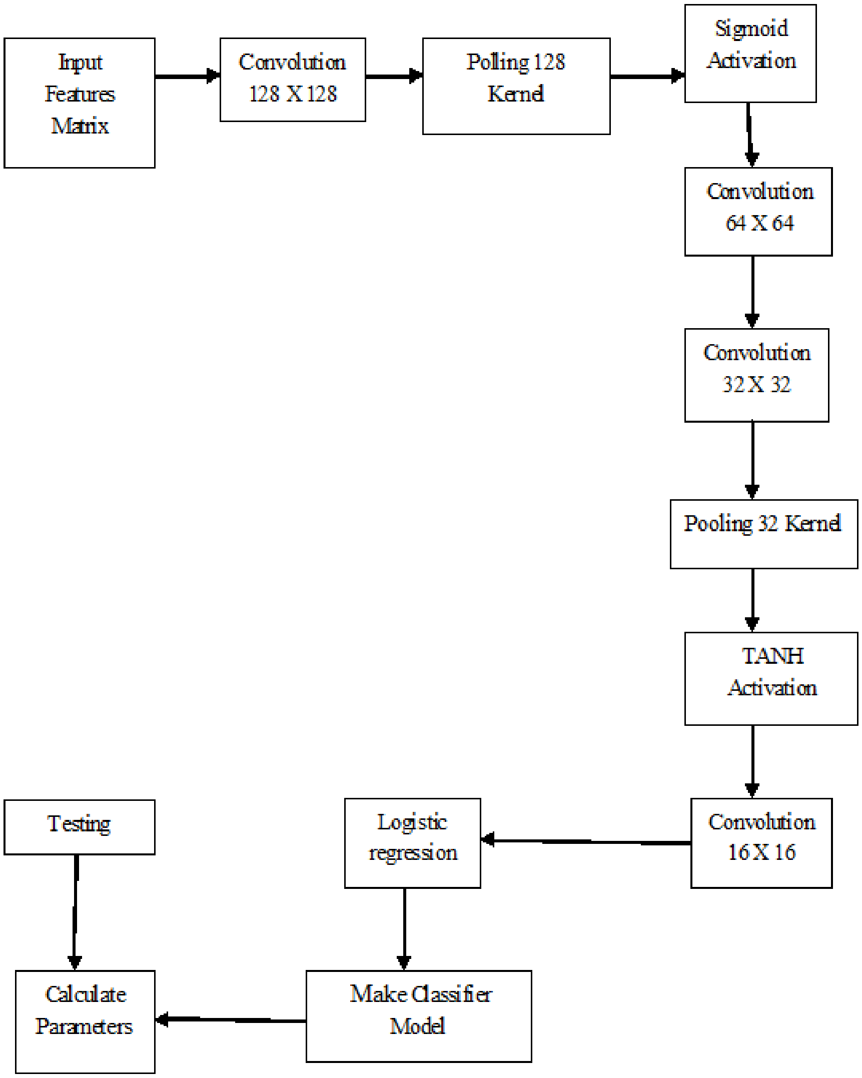


Figure 4.5: Flowchart of Proposed CNN-LR Algorithm

to root words by stemming then gets word vectors are obtained.

Algorithm 4.2 :Features Extraction

Input: Vector of words
Output: features of vectors with class level.

```
While (Vectors of text > 0)
Begin
While (words > 0)
Begin
$F_i$ = frequency of unique word
$f_i$ = log N/ df ......................(1)
$f_i$ = Inverse document frequency.
TF-IDF $\sum_{i=0}^{N} Fi * fi$.......................(2)
Calculate n-gram vector (n=2,3)
Xi = $\sum\lambda=0^{N}$ P($X_i - 1/X_i$) ..........................(3)
$X_i$ = n gram vector
features = $\sum i=0^{N}$ $X_i$ + TF - $TDF_i$ ..................(4)
END
END
```

Both machine learning and deep learning approaches are used in Algorithm 4.2 pseudo code of function extraction. Two steps are used in feature extraction. One is the frequency base feature and document word base features by eqn. 1 and eqn. 2 . This method is called TF-IDF. In other part, n-grams features from eqn. 3 are combined by eqn. 4 to get feature vector which is used for learning of text in machine learning and deep learning approach.

Algorithm 4.3: Data Classification(CNN-LR)

```
Input: feature vector with class label
Output: Learning model for text classification
While (Number of Rows (i) $>$ 0)
Start
While (Number of column (j) $>$0)
Start
Perform Convolution $X_i$
$X_i$ = -y-$a^{(nl)}$ .f' ($z^{(n)}$)........................(5)
$X_i$ = Convolution of i Layer
y = features
$a^{(nl)}$ = n text features
f' ($z^{(n)}$) = transpose of features
Perform Polling and Sigmoid mapping
$X_i^{(l)}$ = (($W^{(l)}$)$^T$ $\delta$(l+1)). f' ($z^{(l)}$)..............................(6)
$X_i^{(l)}$ = Sigmoid mapping of i layer l instances
$W^{(l)}$ = Weight of l instances
(l+1)= Partial differentiation
$z^{(l)}$ = Bias Value
Compute features
$X_i$. $X_i^{(l)}$= (l+1). ($a^{(l)}$)....................................(7)
Learn logistic refression
Learn $X_i$. $X_i^{(l)}$ by loss function
$f_{LR}^{(W)}$ = log (1+e) ..(8)
$f_{LR}^{(W)}$ = logistic function of w features
$y_i$= $i^{th}$ Instances
$X_i$ = $X_i$ Text
w= weight of layer
Stop
```

The pseudo-code of the proposed solution using the CNN-LR presented in Algorithm 4.3.Tweet text is the algorithm's input, which is first preprocessed and then extracted from the function. The first step is to reduce the non-linearity after this learning component begins by convolution, pooling, and activation function. Logistic regression is then used by the Equation. 8. In different EPOCH numbers, the loss and accuracy are then evaluated, which iteratively increases the accuracy and decreases the loss. Equation 7 of Algorithm 4.3 shows the non-linearity of features.

4.2.6 Performance Calculation of an Algorithm

To validate the performance of any system, there are several factors to be considered. The factors could be like time, space, cost, and accuracy. As far the sentiment analysis is considered, classification accuracy is the desired metric of performance. To find the accuracy rate, a confusion matrix should be derived.

Confusion matrix

It is a visual performance assessment of a classified result, given in the form of a matrix. TP When the tweet is positive and the system classifies 'positive'

TN When the tweet is negative and the system classifies 'negative'

FP When the tweet is negative and the system classifies 'positive'

FN When the tweet is positive and the system classifies 'negative'

Table 4.2 shows confusion matrix as mentioned below.

Table 4.2: Confusion Matrix

	Predicted Negative Tweets	**Predicted Positive Tweets**
Observed Negative Tweets	TN	FP
Observed Positive Tweets	FN	TP

Accuracy calculation

The accuracy rate is the measure of how accurately the system classifies the given tweets. The more the accuracy rate, the better is the system. The accuracy calculation equation is given below. The total number of positive tweets that have been identified and the total number of negative tweets that have been identified are divided by the total number of tweets. The error rate is the number of tweets that are incorrectly classified by the classifier. It is the amount of the total number of erroneously accepted tweets as positive and negative, divided by the total number of tweets.

Accurately classified sentences

$$\text{Accuracy} = (\text{TP} + \text{TN})/\text{total no of tweets} \tag{4.2}$$

Wrongly classified sentences

$$\text{Error Rate} = (\text{FP} + \text{FN})/\text{total no of tweets} \tag{4.3}$$

4.3 Novel Security Scheme of Encryption

The minimum policies needed to ensure data security include a strong data protection encryption mechanism and strong authentication mechanisms to ensure data integrity and to prevent unauthorised access. The current study sets its trend of new, safe, constructive approaches to ensure data protection. A study showed that the required keys for data encryption were created by several cloud applications.After data encryption, the MAC code is generated and applied to the encrypted data to be transmitted along with it. A detailed analysis of the prevalent literature revealed that over 50%of cloud service providers adopted a single approach across the entire data. But a situation can occur in which the owner is interested in a small portion of all of his data at a time and even then, his full encrypted data is forwarded by the cloud mechanism. This is degrading efficiency with higher communication costs and rising application costs. Cloud data partitioning is used for effective data storage and for accessing the dominant approach adopted by many cloud providers. The owner's entire data is split into small units, and each unit is encrypted and signed in order to improve data security. Works have shown that the data partition technique increases the storage efficacy of the cloud and error-free data recovery. The owner's whole data is not of equal importance to him. There may be cases where a common portion of data is queried on a daily basis, and some portions of the data may be less queried and sometimes not queried at all. This is also sufficient in cases where single user information is given priority as 1) private, which is the owner's sensitive data and needs to be covered by strict security measures; (2) the public to which any approved user may have access, and in which case the strength of the security measures may be decreased. In cloud scenarios with prioritised data categorizations, it is easier to have various security policies operating at top and priority levels. The need for operational development has made the business outsource its storage and computing requirements.

Works described in [122] demonstrated the efficiency of multiple symmetric key data encryption systems and showed that AES has the highest capabilities for data security. But the drawback is that the data owner had trouble evaluating their data in the encrypted form.Since data owners are naive to encryption schemes, they believe that massive data encryption results in data loss if not properly decrypted, and thus many owners choose not to encrypt schemes when storing their data in the cloud. But there could be threat to the unencrypted data in the cloud.

4.3.1 The Proposed Method ECDH-SAHE

This study focused on the security concerns of big data and considered the fair use of cloud computing methodology. The suggested solution, ECDH-SAHE, aimed to improve safety assurance efficiency. The ECDH algorithm was developed in conjunction with the key algorithm supporting the SAHE model to dynamically replace the data-based packages for encryption under different time-based constraints. The test evaluations demonstrated a superior and adaptive output for the proposed strategy [52].

Algorithm 4.4: Encryption Algorithm(ECDH-SAHE)

```
Algorithm (Encryption)
Input: Text
Output: Encrypted Text and Key
1. Start
2. Elicptic curve [EC] (Text)
3. Key=EC (Text)
4. Client=DH(Key)
5. Str=String (Text)
6. Gen_bin=binary(str)
7. s=0, i= length (Gen_bin)
8. while (i is not zero) do
9. x= i mod 10
10. y= x mod 10 and i/10
11. s=x+y
12. end while
13. $T_{text}$= left_shift (str, s)
14. Stop
```

The flowchart of an implementation of the algorithm 4.4 is given in Figure 4.7.

Algorithm 4.5: Decryption Algorithm(ECDH-SAHD)

Input: Key and Encrypted Text

Output: Original Text

```
Start
L=32
i= $|key/L|$
s=0
while (i is not zero) do
    x= i mod 10
    y=x mod 10 and i/10
    s = x+y
End while
z= right_shift ($T_{Text}$, Sum)
$O_{Text}$= $(z)^c$
$O_{Tex}$t= decimal ($O_{Text}$)
Stop
```

The flowchart of an implementation of the algorithm 4.5 is given in Figure 4.8.

This research work proposes capable encryption, namely SAHE i.e. Shifted Adaption Homomomorphism Encryption, which is considered by all current research studies to be the better alternative.

4.3.2 Steps of implementing proposed technique of ECDH-SAHE

Four processes are included in SAHE (Shifted Adaption Homomorphism Encryption), namely Generate Key, Encrypt, Decrypt, and Implement. The 'Length' parameter represents the length (bit length) of integers that are considered in the schema given. It denotes the final size of the hidden key that is generated along with the primary key components.The added noise is utilized as a secret key and is produced by implementing the functions such as: R S or L S using Encrypt for given set of integers. Flowchart of key generation is shown in Figure 4.6.

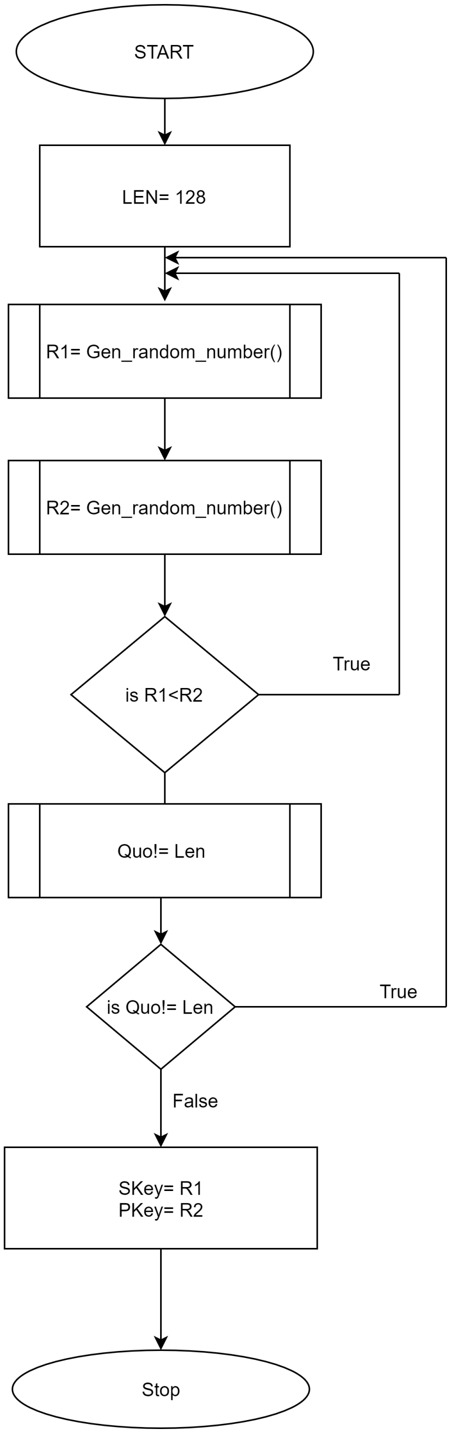


Figure 4.6: Proposed Flowchart of Generating Key

Generate Key (Length)

Elliptic Curve Diffie Hellman (ECDH) is a key exchange protocol that is considered to be secure for larger bit lengths. ECDH allows two communicants to exchange a shared secret over an insecure medium [104]. This shared key or public key is used to derive another key namely session key that is used for encryption. This will initially consider the Len parameter as an input that refers to the length of the constant. The Len value is taken from the schema, and then a random Len-bit decimal number is generated, called R 1. In this study, we view the bit of variable length as an appropriate length to infuse the secret key. Make another random Len-bit decimal number, say R 2, where R 1 > R 2 and then calculate Quo = (Round(R 1 /R 2)); now compare Quo to Len or not; if true, go ahead; otherwise, recreate the values of R 1 and R 2. Upon obtaining the Len, return the R 1 as the SKey secret key and R 2 as the PKey public key.

Encrypt (str)

This function uses str (a set of alphanumeric characters also referred to as string) as the schema input. This string str is inverted with the M^C(str) inversion function and P_{Key} is used to execute the L_S left-shift function on the inverted str. Finally, the cypher text C_{Text} of size Len x 2 is generated by this technique. This implementation, therefore, gives:

$$\mathrm{C_{Text} = L_5\left(M^C(str), sum\left(P_{Key}\right)\right)}$$

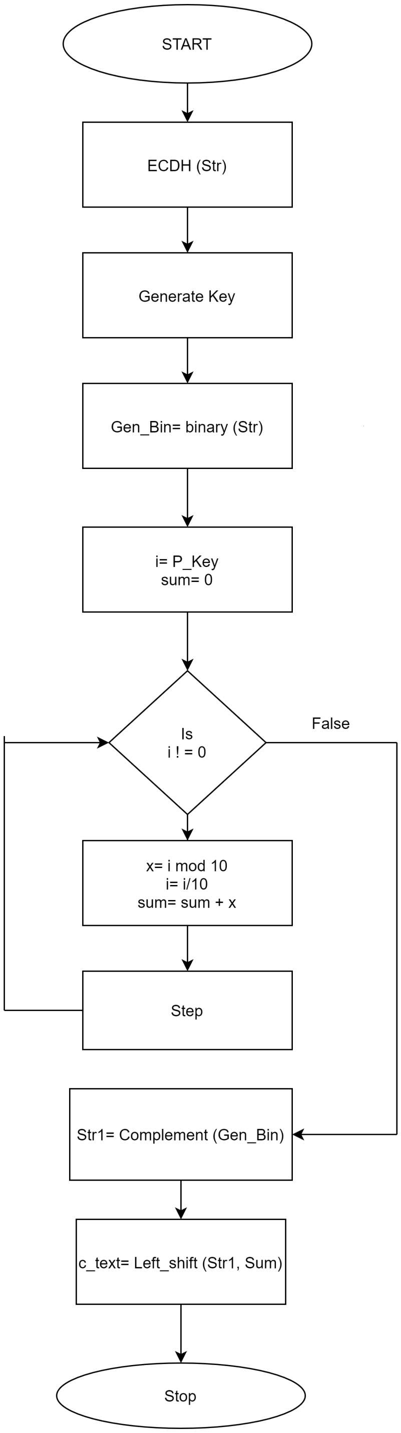


Figure 4.7: Proposed Flowchart of ECDH-SAHE

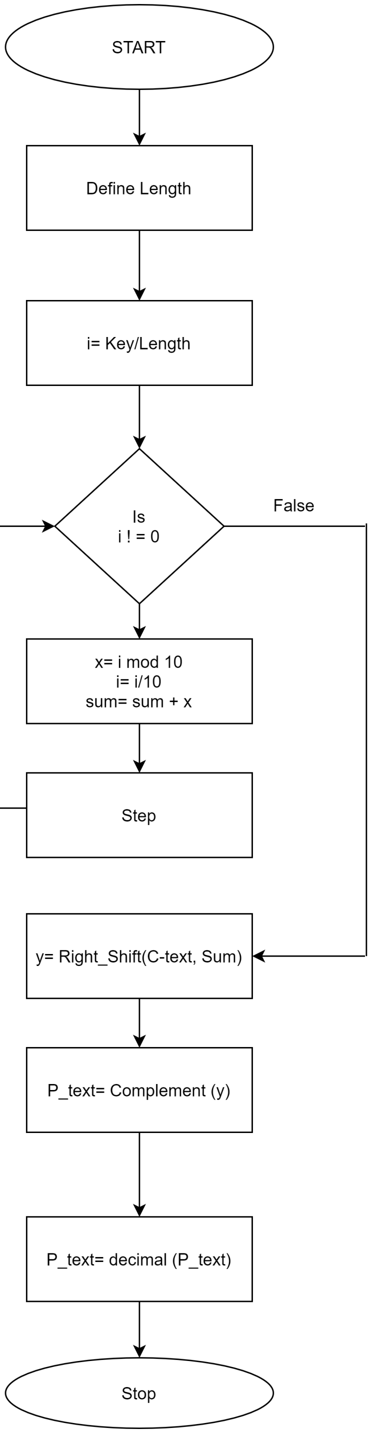


Figure 4.8: Proposed Flowchart of ECDH-SAHD

Decrypt (C_{Text})

As an input variable, this function takes C_{Text}. Sum (S_{key}/Len) is first computed as part of the decryption process, then the right shift R_S function is executed on the cypher text. The resulting text is given to the M^C inversion function, and now the generated string is P_{Text} plain text, and the function executed to obtain plain text is:

$$\mathrm{P_{Text}} = \mathrm{M^C}\left(\mathrm{R_5}\left(\mathrm{C_{Text}}, \mathrm{sum}\left(\mathrm{S_{key}}/\mathrm{Len}\right)\right)\right)$$

Implement (P_{Text})

Plain text is sent to this function as input. In order to evaluate the data with XOR & AND gates, the generated cypher text is passed to a binary circuit where the multiplication and addition operations are executed mathematically on values of A that may be real or integer values. The P_{Key} size is O(n^7). On addition or multiplication, SAHE depends. SAHE (Shifted Adaption Homomorphism Encryption) adapts and eliminates noise in the implementation function very quickly.

As clearly discussed above, it is a major issue that the real time data is a combination of important and not-so important. Therefore they need different treatments while encrypting in order to save time and space.

The proposed research work is an optimal solution to classify the given text into sensitive and non-sensitive. In the first part of the research the accuracy of various classification algorithms is compared and the best one is chosen for the further process. Further, only the sensitive data is encrypted so as to save time , space and storage. The detailed contribution is discussed in subsequent chapters.

CHAPTER 5

RESULTS AND DISCUSSION

Outline

This chapter discusses the results formulated in research work. A comparative analysis is used as a tool. Sentiment analysis is a technique of classifying the given text for its polarity. This classification is efficient and helps in decision making. The data grows fast in size and as a result, it becomes voluminous and at this point, the task of classification becomes tedious and unmanageable. To overcome this limitation an automated tool, technique, or better algorithm is required with the aim of saving valuable time, space, or accuracy. This research work is efficient in classifying the text into sensitive or non-sensitive to gain accuracy. The sensitive data will be encrypted in less time and using less space. Many algorithms are available for this purpose, so a comparative study is conducted and then a novel model is proposed for better results.

5.1 Description of the Dataset

Many websites provide access to a variety of datasets suitable for data mining and machine learning experimentation. The dataset taken for this research purpose is Twitter. Datasets consist of Tweets and it was collected from Twitter by using Twitter API. A search query was sent to Twitter and it resulted in JSON format. The dataset consists of 1 Lakh Tweets. http://twitter.com

Following are the parameters that are considered while studying an algorithm. They are time, space, memory, and accuracy. This research aims at studying the accuracy rate of different algorithms and is efficient in proposing an entirely new method that follows the CNN-LR approach. The use of the CNN-LR approach is accurate and show a better rate, which is discussed in detail in the research undertaken. The proposed novel method is highly recommended through the conducted research for producing accurate results.

The implementation phase is performed in the following two steps. They are the unigram approach and the novel method. Though there are many social websites available in the market today, Twitter is the predominant of all that uses a short message of 140 characters. This characteristic is best suited and followed in the research with the aim of analyzing or classifying the text for determining polarity. The tweets form the basis for the datasets. The dataset ranges from one thousand to over one lakhs tweets.

The following research work is also illustrative of comparison amongst machine learning and deep learning approach. In the proposed approach there is a hybridization of machine learning and deep learning approach. In the proposed approach CNN and the logistic regression hybrid approach have been used. In existing approaches KNN (non-parametric approach), SVM (Support Vector Machines), and Decision Trees and Neural Networks are used. As input, tweet the chosen dataset as a text has been taken. Tweets are a mixture of tweets that are gathered with the aid of REST API. In the dataset 100,000 tweets for training and 10000 tweets for testing have been collected.

5.2 Accuracy rate of various classification algorithms

This research is carried out in two phases, the first phase is a comparative study of various algorithms, and the second phase is suggesting and implementing a novel method using the CNN-LR approach. The comparative study of various algorithms is carried out by using PYTHON. This tool is enriched with many features for implementing analysis using different classification algorithms. A visual environment is available for displaying the result in a better way. PYTHON language is included for programming. A comparative analysis is studied using various datasets. The dataset for comparison is chosen and four algorithms namely KNN, SVM, Neural, and CNN are chosen for the accuracy comparison. This is better illustrates in Table 5.1 that the CNN-LR algorithm gives more accurate results for the given dataset.

Table 5.1: Accuracy, Precision, Recall of various algorithms in percentage

Algorithm	**Accuracy**	**Precision**	**Recall**
KNN	77.28	74.34	72.12
SVM	78.53	76.12	73.12
Decision Tree	68.65	69.12	70
Neural Network	77.51	76.13	75.23
Proposed(CNN-logistic)	79.8	80.23	81.34

5.2.1 Reason for Selecting CNN-LR

On conducting experimental analyses and following KNN to Neural Network, the improvement in results is seen. This is better illustrates in Table 5.1 and these results further encourage the choice of the layered convolution network.

- In machine learning approaches, features depend on linear structures and there is no non-linearity.
- In machine learning, classification is optimized by activation function.
- In CNN, non-linearity is highly effective and reduces the latent features.

The reasons cited above motivate for use of non-linearity base features that are based upon convolution approach and learning by machine learning approach such as logistic regression. Logistic regression shows the same behaviour of activation function like Sigmoid function and TANH function.

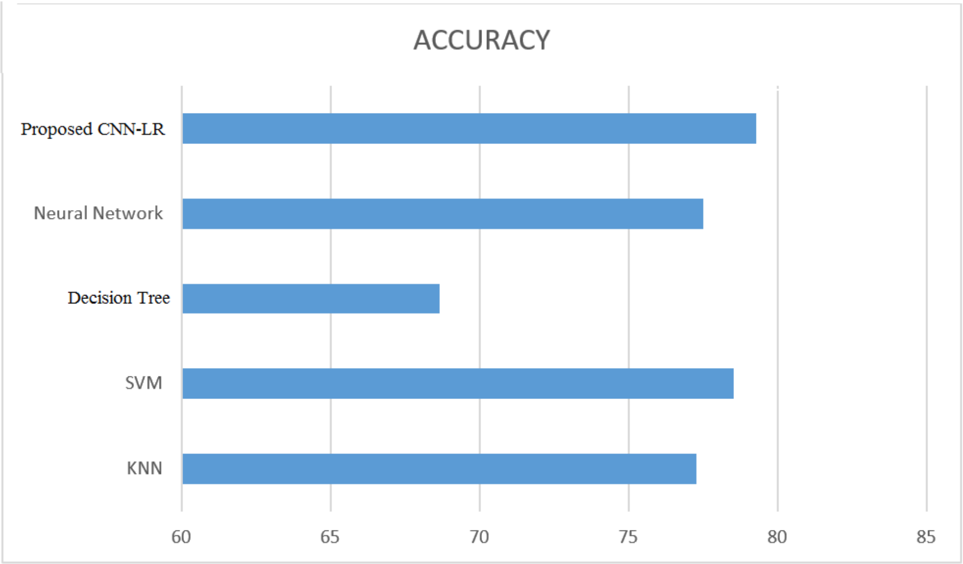


Figure 5.1: Comparison of accuracy of different classifier

5.2.2 Reason for Improving Performance of Proposed approach

It improves the behaviour of feature extraction by a layered convolution-based approach that is based upon the convoluted feature and mapping it in an abstract way that is further responsible for the non-linearity.

Figure 5.2 depicts the different Epochs and improves the accuracy of CNN-Logistic but not as effective as a Neural Network. Because the CNN-Logistic approach uses non-linear features with linear features but Neural Network uses linear features and ignores the non-linear features. Non-linearity comes from the CNN approach with Sigmoid and TANH function that induces the learning most effectively.

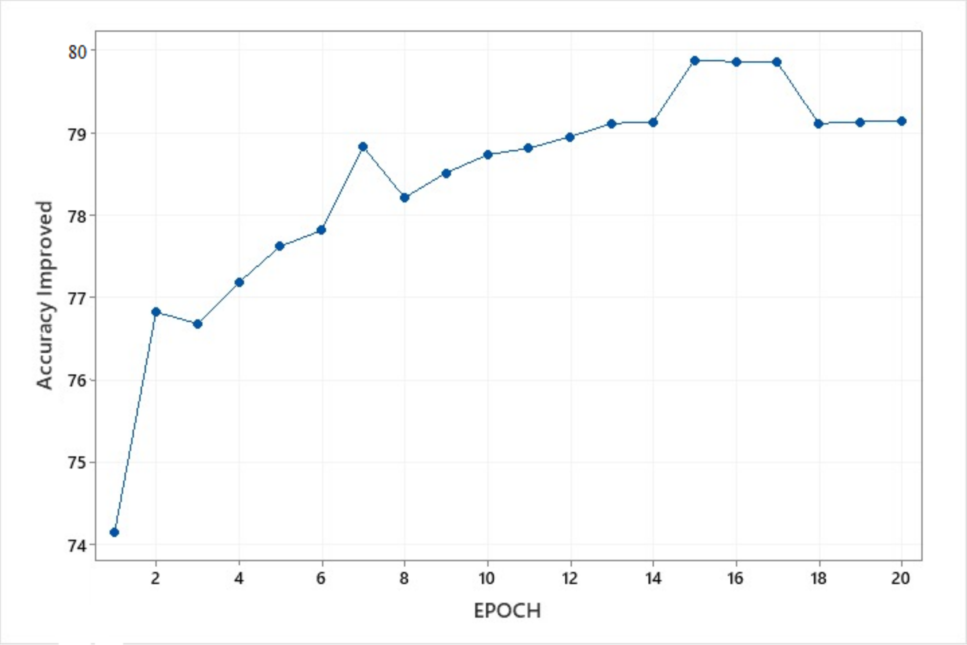


Figure 5.2: Comparison of different EPOCH in CNN-LR proposed approach

5.3 Result of Various Encryption Techniques

The below figures show the comparison of existing and the proposed approach based on encryption time, decryption time and storage space. The results obtained show that the proposed approach (ECDH-SAHE) reduce time and storage space as required.

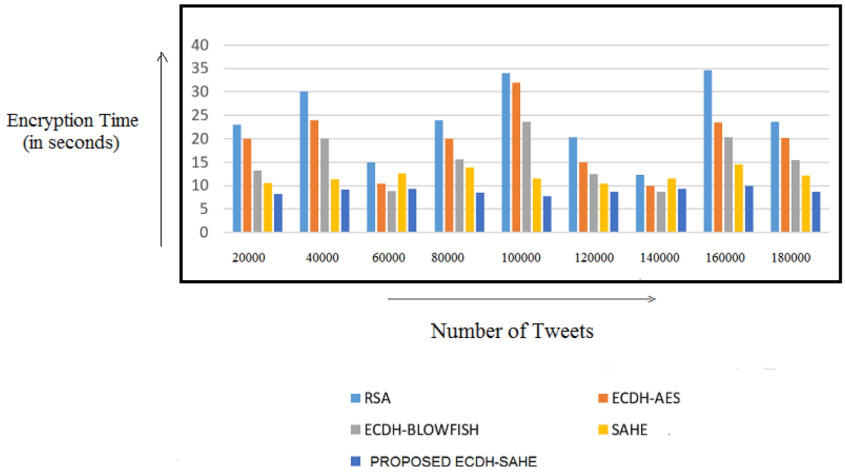


Figure 5.3: Comparison of the proposed and existing encryption time approach based on different dataset sizes

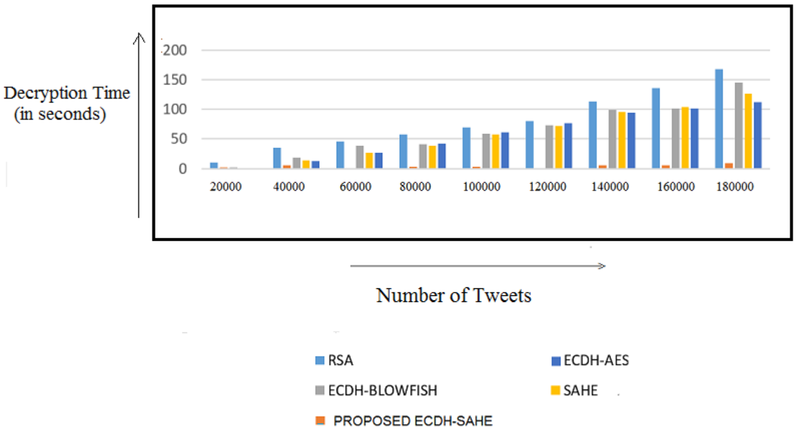


Figure 5.4: Comparison of proposed and existing approach decryption time based on different size dataset

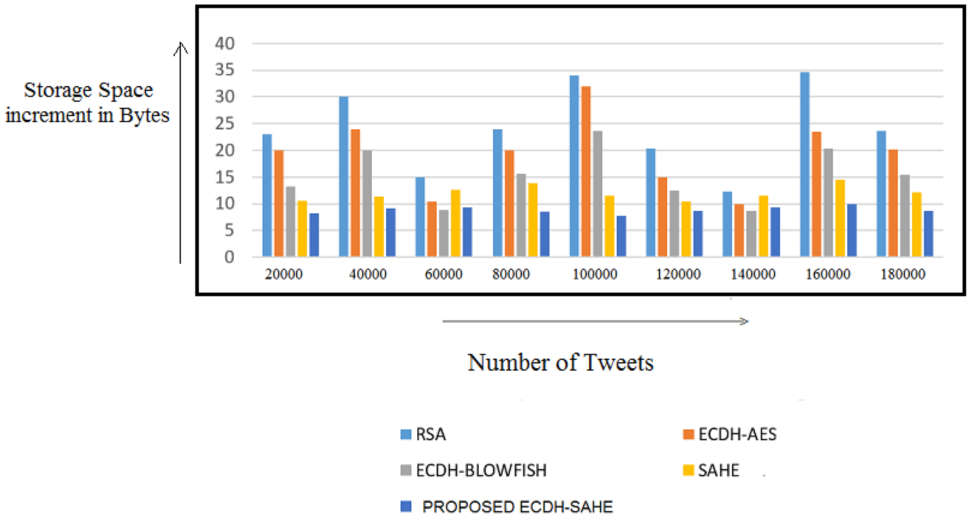


Figure 5.5: Comparison of proposed and existing approach in terms of storage space

5.3.1 Complexity Analysis

The performance of the proposed and existing approach is highlighted in the research that is based on different sizes of the dataset. The graphs obtained demonstrate that the proposed methodology significantly reduces encryption, decryption time, and storage. If we compare the performances of a small and large dataset, we can observe that an increase in size only accommodates overhead in an effective manner. Thus storage and time do not increase much with the increase in the size of data. This work provides great motivation for the large dataset and effectively checks the performance of the proposed approach. While encrypting data, encryption algorithms use a considerable amount of Processor time and space at the time of encryption. The purpose of this research is to compare and find space for the various encryption algorithms. The complexity of encrypted and decrypted data could be determined by using the encryption algorithms complexities. In this research work, the five most commonly used algorithms are compared against each other.

5.3.2 Time and Space Complexity Analysis of various Encryption algorithms

This research work examines the fundamental issue of data security in the context of cloud computing mechanisms. Concerning cloud computing methodology, we present the model of data security dependent on the investigation of the architecture of the cloud. This will concentrate on the protection issues of big data and thought about the reasonable usage in the methodology of cloud computing. The proposed approach, ECDH-SAHE, was intended to augment the productivity of security assurances. ECDH algorithm in conjunction with the main algorithm supporting the SAHE model was created to dynamically substitute the data-based packages for encryptions under various time-based constraints.

This research work provides a comparison between the various symmetric encryption algorithms such as RSA, AES, BLOWFISH, SAHE Algorithm. The study is based on the algorithms' architecture, the aspects of security, and the limitations they have. The comparison between different algorithms show that asymmetric algorithms are superior in protection, but processing takes more time and requires more memory. In general, for the key exchange process, asymmetric algorithms are used and symmetric algorithms are used for the process of encryption and decryption[127]. The results show that if time and space are a major factor in the application, the most suitable algorithm is the proposed ECDH-SAHE. Compared to other algorithms, ECDH-SAHE gains less space and less time after comparing all algorithms. Space and time complexity is also compared with other algorithms such as RSA, AES, BLOWFISH.

It can be shown that ECDH-SAHE typically performs better than other algorithms. The improvement in the proposed approach is seen for the following reasons:

- It reduces time overhead using a binary stream instead of a unary stream of text.

- Reduces the storage space by shifting left at the time of encryption and right during decryption.

- Analysis of the shifting generally depends on the binary streaming and indirectly depends upon the security-based enhancement. So, the proposed approach improves the security and time-based overhead.

- It also reduces storage because of the homomorphism nature of selecting bitstream.

5.4 Summary

The parameters that were considered while studying an algorithm were time, space, memory, and accuracy. In the proposed approach there is a hybridization of machine learning and deep learning approach. In the proposed approach CNN and the logistic regression hybrid approach have been used. This research is carried out in two phases, the first phase is a comparative study of various algorithms, and the second phase is suggesting and implementing a novel method using the CNN-LR approach. Because the CNN-Logistic approach uses non-linear features with linear features but Neural Network uses linear features and ignores the non-linear features. Non-linearity comes from the CNN approach with Sigmoid and TANH function that induces the learning most effectively.

The results obtained show that the proposed approach (ECDH-SAHE) reduces time and storage space as required. Concerning cloud computing methodology, we present the model of data security dependent on the investigation of the architecture of the cloud. The results show that if time and space are a major factor in the application, the most suitable algorithm is the proposed ECDH-SAHE.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

The goal of this research is to develop an efficient algorithm that has potential of classifying big data into sensitive or non-sensitive parts for a better result. The efficiency is measured in terms of the accuracy rate of classification of the given document. The novel approach that is developed is based on CNN-LR. The text for classification is taken from Twitter. But recommended novel approach could be applied to any text data. The classification of the text on the basis of its polarity can take different forms that are positive, negative, and neutral. The tweets are classified as negative tweets and positive tweets based on the statistical model, the CNN-LR approach.

The proposed novel method classifies tweets as positive and negative. The remaining tweets, which the system is not able to classify is considered neutral. There are 2- class classifiers and the neutral value identified is not a class or polarities that are used as identifiers.

There are certain limitations to this research work. Though this implementation could be applied to any dataset, this work is limited to only tweets as datasets. There are several social media networking sites available with plenty of datasets. When applied for domain-specific data, more insights could be retrieved. The text that is considered for the purpose of review can take up the form of sentence, paragraphs, or a document. Since this work is based on a statistical model, the time taken to classify is faster.

The second limitation in this work is, pre-processing. The pre- processing i.e.data cleaning is done manually. As mentioned in the previous chapters, the tweets from Twitter contain some unwanted data like emoticons, abbreviations, other language alphabets, URLs, and sarcasm. These unwanted details are removed manually and stored in a file.

This research work is based on a statistical model, but sentiment analysis involves emotions, sentiments, voice, intelligence, and opinion of individuals. The taste and value system of one person may not be the same. All these factors depend on person to person, so the comment, review, and feedback given by users should also consider those factors so that accurate results are claimed. This involves Natural Language Processing technqiues; that is often not considered in this research work. In this research, sentiment classification will not be able to deploy in mobile phones or web devices, despite the fact that they are the perennial source of tweets.

Secondly, this thesis has aimed to reduce the storage spaces and time as well as providing security for user sensitive data through a security model and implementation of improved security algorithm. The storage performance is very important because it affects application performance and slows down the execution process. This study has provided answers to the two research questions raised in the conducted research work. The first research question on identifying the existing security techniques and their efficiency for enhancing storage spaces of encrypted data that is effectively presented. Their efficiencies in terms of benefits and limitations were studied that aided in determining the existing gap. This further leads to the second research question, that based upon improved security algorithm that is developed to address the security issues while enhancing the storage spaces and time. This research work was able to design a novel encrypted algorithm ECDH-SAHE for securing sensitive data. The comparison between various techniques is done using encryption time, decryption time, and throughput as performance metrics. The study results were compared with existing similar works. Our encryption time and decryption time are faster than other existing techniques. Hence, the results are very encouraging as well as effective.

6.2 Future Enhancement

This research suggests a novel method using the CNN-LR approach for the classification of tweets more accurately. A milestone has been crossed in this effort. This research will open the way for avenues in various fields of research. In the future, research will be continued to incorporate Artificial Intelligence (AI) and the Internet of Things (IoT). A new field of engineering is growing that will combine machine learning with AI for better analyses over big data with much accurate prediction. When this engineering is applied to IoT, there will be more promising results expected. This research can be continued to obtain automatic analyzer solutions for smart devices. The IoT (Internet of Things) is our generation's future and it is not easy to handle various types of data together. Because of its simple installation and widespread applications in big data analysis, Splunk has gained immense popularity. As the IT industry strengthens its arms on a regular basis, Splunk's scope grows rapidly. Each IT business, large or small, needs to handle the data of computer, and Splunk is undoubtedly the best to do so in the market. It doesn't end there, it still adds to infrastructure more features, making it easier to use. With its current growth pace, rivalry will soon be influenced by its competitors. So in future, Splunk is recommended for performing visualization and pattern recognition. This research is limited to the 2-class classifier, in the future this could be extended to a multi-class classifier. This research determines only the polarity of positive or negative. This could be extended to add contextual semantics of the content of the text and better reveals of the emotion and feeling of the speaker.

In the future, more attractive and impressive visualization techniques could be implemented. Further, the dataset used here is the tweets of sentences, in the future; this could be extended to paragraphs and documents. This research is confined to web data, whereas in the future, this could be extended to mobile data. Sentiment analysis is growing in different depth and width. Deep sentiment analysis of the text is a natural language processing problem where the given text is understood and underlying meaning and semantics are predicted. In such a scenario, the overall text classification of positive or negative alone is not enough. The research could be extended to know the separate topics spoken in the text. In future, the research is further carried using Natural Language Processing technique over using machine learning techniques.

For future research directions; Known cryptanalytic attacks such as key recovery attack, side channel attack and the known-plaintext attack would be tested on the current implemented work. The research work is evaluated using time and space performance. There is a need to do future exploration considering other performance metrics like memory usage and implementation on Advanced Reduced instruction set computer Machine (ARM) architecture with reduced logic gates, clock cycle, and data path.

Bibliography

[1] Abdul-Mageed, M., Diab, M., and Korayem, M. (2011). Subjectivity and sentiment analysis of modern standard arabic. pages 587–591.

[2] Abed, F. (2013). A proposed method of information hiding based on hybrid cryptography and steganography dr.

[3] Acar, T., Belenkiy, M., Nguyen, L., and Ellison, C. (2011). Key management in distributed systems.

[4] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R. J. (2011). Sentiment analysis of twitter data. In *Proceedings of the workshop on language in social media (LSM 2011)*, pages 30–38.

[5] Ahmed, J. and Ahmed, M. (2020). A framework for sentiment analysis of online news articles. 11:267–274.

[6] Almorsy, M., Grundy, J., and Ibrahim, A. S. (2011). Collaboration-based cloud computing security management framework. In *2011 IEEE 4th International Conference on Cloud Computing*, pages 364–371. IEEE.

[7] Almulla, S. and Yeun, C. (2010). Cloud computing security management. pages 1 – 7.

[8] Arras, L., Montavon, G., Müller, K.-R., and Samek, W. (2017). Explaining recurrent neural network predictions in sentiment analysis. *arXiv preprint arXiv:1706.07206*.

[9] Aung, S. S., Nagayama, I., and Tamaki, S. (2016). Intelligent traffic prediction by multi-sensor fusion using multi-threaded machine learning. *IEIE Transactions on Smart Processing & Computing*, 5(6):430–439.

[10] Aung, S. S. and Naing, T. T. (2015). Naïve bayes classifier based traffic prediction system on cloud infrastructure. In *2015 6th International Conference on Intelligent Systems, Modelling and Simulation*, pages 193–198. IEEE.

[11] Balasubramanian, R. and Aramuthan (2021). Security problems and possible security approaches in cloud computing.

[12] Bao, Y. and Ishii, N. (2002). Combining multiple k-nearest neighbor classifiers for text classification by reducts. pages 340–347.

[13] Bao, Y., Ishii, N., and Du, X. (2004). Combining multiple k-nearest neighbor classifiers using different distance functions. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 634–641. Springer.

[14] Barbosa, L. and Feng, J. (2010). Robust sentiment detection on twitter from biased and noisy data. volume 2, pages 36–44.

[15] Behl, A. and Behl, K. (2012). An analysis of cloud computing security issues. pages 109–114.

[16] Beineke, P., Hastie, T., and Vaithyanathan, S. (2004). The sentimental factor: Improving review classification via human provided information. pages 263–270.

[17] Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., and Jurafsky, D. (2006). Extracting opinion propositions and opinion holders using syntactic and lexical cues. In *Computing Attitude and Affect in Text: Theory and Applications*, pages 125–141. Springer.

[18] Bhonde, R., Bhagwat, B., Ingulkar, S., and Pande, A. (2015). Sentiment analysis based on dictionary approach. *International Journal of Emerging Engineering Research and Technology*, 3(1):51–55.

[19] Bifet, A. and Frank, E. (2010). Sentiment knowledge discovery in twitter streaming data. In *International conference on discovery science*, pages 1–15. Springer.

[20] Brücher, H., Knolmayer, G., and Mittermayer, M.-A. (2002). Document classification methods for organizing explicit knowledge.

[21] Cane, D., Hirschman, D., Speare, P., and Vaitzblit, L. (1999). Secure file archive through encryption key management. US Patent 5,940,507.

[22] Cao, D., Ji, R., Lin, D., and Li, S. (2016). Visual sentiment topic model based microblog image sentiment analysis. *Multimedia Tools and Applications*, 75(15):8955–8968.

[23] Challal, Y. and Seba, H. (2005). Group key management protocols: A novel taxonomy. *Information Technology - IT*, 2.

[24] Chan, H., Lacka, E., Yee, R., and Lim, M. (2014). A case study on mining social media data. *2014 IEEE International Conference on Industrial Engineering and Engineering Management*, pages 593–596.

[25] Chauhan, A. and Gupta, J. (2017). A novel technique of cloud security based on hybrid encryption by blowfish and md5. In *2017 4th International conference on signal processing, computing and control (ISPCC)*, pages 349–355. IEEE.

[26] Chen, L.-S., Liu, C.-H., and Chiu, H.-J. (2011). A neural network based approach for sentiment classification in the blogosphere. *Journal of Informetrics*, 5(2):313–322.

[27] Chen, X., Madhavan, K., and Vorvoreanu, M. (2013). A web-based tool for collaborative social media data analysis. In *2013 International Conference on Cloud and Green Computing*, pages 383–388. IEEE.

[28] Chou, K.-Y., Chen, Y.-R., and Tzeng, W.-G. (2011). An efficient and secure group key management scheme supporting frequent key updates on pay-tv systems. pages 1–8.

[29] Churi, P., Kalelkar, M., and Save, B. (2014). Jsh algorithm: A password encryption technique using jumbling-salting-hashing. *International Journal of Computer Applications*, 92.

[30] Conejero, J., Burnap, P., Rana, O., and Morgan, J. (2013). Scaling archived social media data analysis using a hadoop cloud.

[31] D'Andrea, A., Ferri, F., Grifoni, P., and Guzzo, T. (2015). Approaches, tools and applications for sentiment analysis implementation. *International Journal of Computer Applications*, 125:26–33.

[32] Darwazeh, N. S., Al-Qassas, R. S., AlDosari, F., et al. (2015). A secure cloud computing model based on data classification. *Procedia Computer Science*, 52:1153–1158.

[33] Das, S. R. and Chen, M. Y. (2007). Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management science*, 53(9):1375–1388.

[34] Dave, K., Lawrence, S., and Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528.

[35] Davidov, D., Tsur, O., and Rappoport, A. (2010). Enhanced sentiment learning using twitter hashtags and smileys. volume 2, pages 241–249.

[36] Dhande, L. L. and Patnaik, G. (2014). Review of sentiment analysis using naive bayes and neural network classifier. *International Journal of Scientific Engineering and Technology Research (IJSETR)*, 3(7):1110–1113.

[37] Ding, X., Liu, B., and Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, pages 231–240.

[38] Diwan, V., Malhotra, S., and Jain, R. (2014). Cloud security solutions: Comparison among various cryptographic algorithms. *IJARCSSE, April.*

[39] Domingos, P. and Pazzani, M. (1998). On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29.

[40] Ediger, D., Jiang, K., Riedy, J., Bader, D., Corley, C., Farber, R., and Reynolds, W. (2010). Massive social network analysis: Mining twitter for social good. volume 0, pages 583–593.

[41] Ehrsam, W. F., Matyas, S. M., Meyer, C. H., and Tuchman, W. L. (1978). A cryptographic key management scheme for implementing the data encryption standard. *IBM Systems Journal*, 17(2):106–125.

[42] Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification [j]. *Journal of Machine Learning Research - JMLR*, 3.

[43] Gamon, M. (2004). Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 841–847.

[44] Gao, J., Galley, M., and Li, L. (2018). Neural approaches to conversational ai. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1371–1374.

[45] Gardner, M. W. and Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–2636.

[46] Gitanjali, K. L. (2019). A novel approach of sensitive data classification using convolution neural network and logistic regression.

[47] Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12).

[48] Govindarajan, M. (2014). Sentiment classification of movie reviews using hybrid method. *International Journal of Advances in Science Engineering and Technology*, 1(3):73–77.

[49] Grigoriev, D. and Ponomarenko, I. (2006). Homomorphic public-key cryptosystems and encrypting boolean circuits. *Applicable Algebra in Engin., Communic. and Comput.*, 17:239–255.

[50] Gupta, D. S. (2012). Securely management crypgraphic keys used within acloud environment. In *NIST Cryptographic Key management workshop*.

[51] Gupta, G. and Lakhwani, K. (2020a). Big data classification techniques: A systematic literature. *Journal of Natural Remedies*, 21(2):S1.

[52] Gupta, G. and Lakhwani, K. (2020b). Improved encryption of big data by shift homomorphic with ecdh approach. *Test Engineering and Management Journal*, 83:25416– 25424.

[53] Hagge, M., von Hoffen, M., Betzing, J., and Becker, J. (2017). Design and implementation of a toolkit for the aspect-based sentiment analysis of tweets. pages 379–387.

[54] Han, Y., Brezany, P., and Janciak, I. (2009). Cloud-enabled scalable decision tree construction. *Semantics, Knowledge and Grid, International Conference on*, 0:128–135.

[55] Hao, M., Rohrdantz, C., Janetzko, H., Dayal, U., Keim, D. A., Haug, L.-E., and Hsu, M.-C. (2011). Visual sentiment analysis on twitter data streams. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 277–278. IEEE.

[56] Hashizume, K., Rosado, D., Fernández-Medina, E., and Fernández, E. (2013). An analysis of security issues for cloud computing. *Journal of Internet Services and Applications*, 4.

[57] Hearst, M. (1999). Direction based text interpretation as an information access refinement.

[58] Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

[59] Hurst, M. and Nigam, K. (2004). Retrieving topical sentiments from online document collections. *Proc. of the Document Recognition and Retrieval XI*, pages 27–34.

[60] Ikonomakis, E., Kotsiantis, S., and Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS transactions on computers*, 4:966–974.

[61] Ikram, A., Ibrahim, S., Sardaraz, M., Tahir, M., Bajwa, H., and Bach, C. (2013). Neural network based cloud computing platform for bioinformatics. pages 1–6.

[62] Ingle, R. and Sivakumar, G. (2007). Tunable group key agreement. pages 1017–1024.

[63] Isa, D., Lee, L. H., Kallimani, V., and Rajkumar, R. (2008). Text document preprocessing with the bayes formula for classification using the support vector machine. *Knowledge and Data Engineering, IEEE Transactions on*, 20:1264–1272.

[64] Isah, H., Trundle, P., and Neagu, D. (2014). Social media analysis for product safety using text mining and sentiment analysis. In *2014 14th UK workshop on computational intelligence (UKCI)*, pages 1–7. IEEE.

[65] Jain, T. and Dipak, N. (2010). Recognizing contextual polarity in phrase-level sentiment analysis. *International Journal of Computer Applications*, 7.

[66] Jang-Jaccard, J., Manraj, A., and Nepal, S. (2012). Portable key management service for cloud storage. pages 147–156.

[67] Ji, R., Cao, D., Zhou, Y., and Chen, F. (2016). Survey of visual sentiment prediction for social media analysis. *Frontiers of Computer Science*, 10.

[68] John Justin, M. and Manimurugan, S. (2012). A survey on various encryption techniques. *International Journal of Soft Computing and Engineering*, 2231:2307.

[69] Joshi, A. and Joshi, B. (2011). A randomized approach for cryptography. pages 293–296.

[70] Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). Bag of tricks for efficient text classification. pages 427–431.

[71] Kalarani, P. and Selva Brunda, S. (2015). An overview on research challenges in opinion mining and sentiment analysis. *Int. J. Innov. Res. Comput. Commun. Eng*, 3(10):1–6.

[72] Kalpana, G., Kumar, P. V., Aljawarneh, S., and Krishnaiah, R.

(2018). Shifted adaption homomorphism encryption for mobile and cloud learning. *Computers and Electrical Engineering*, 65:178–195.

[73] Kaminskas, M. and Ricci, F. (2012). Contextual music information retrieval and recommendation: State of the art and challenges. *Computer Science Review*, 6(2-3):89–119.

[74] Kang, K., Yoon, C., and Kim, E. (2016a). Identifying depressive users in twitter using multimodal analysis. pages 231–238.

[75] Kang, K., Yoon, C., and Kim, E. Y. (2016b). Identifying depressive users in twitter using multimodal analysis. In *2016 International Conference on Big Data and Smart Computing (BigComp)*, pages 231–238. IEEE Computer Society.

[76] Kester, Q.-A. (2013). A hybrid cryptosystem based on vigenere cipher and columnar transposition cipher. *International Journal of Advanced Technology and Engineering Research(IJATER)*, 3.

[77] Kim, H.-J., Kim, H.-I., and Chang, J.-W. (2017). A privacy-preserving knn classification algorithm using yao's garbled circuit on cloud computing. pages 766–769.

[78] Kim, S.-M. and Hovy, E. (2004). Determining the sentiment of opinions. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1367–1373.

[79] Kouloumpis, E., Wilson, T., and Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg!

[80] Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). Imagenet

classification with deep convolutional neural networks. pages 1097–1105.

[81] Lai, S., Xu, L., Liu, K., and Zhao, J. (2015). Recurrent convolutional neural networks for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

[82] Li, C. and Park, S. (2009). An efficient document classification model using an improved back propagation neural network and singular value decomposition. *Expert Syst. Appl.*, 36:3208–3215.

[83] Li, Y., Gai, K., Qiu, L., Qiu, M., and Zhao, H. (2016a). Intelligent cryptography approach for secure distributed big data storage in cloud computing. *Information Sciences*, 387.

[84] Li, Y., Hao, Z., and Lei, H. (2016b). Survey of convolutional neural network. *Journal of Computer Applications*, 36(9):2508–2515.

[85] Liang, M. and Hu, X. (2015). Recurrent convolutional neural network for object recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3367–3375.

[86] Lin, W. and Lee, D. (2012). Traceback attacks in cloud - pebbletrace botnet. pages 417–426.

[87] Liu, B. (2010). Sentiment analysis: A multi-faceted problem. *IEEE Intelligent Systems*, 25(3):76–80.

[88] Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

[89] Liu, B., Blasch, E., Chen, Y., Shen, D., and Chen, G. (2013). Scalable sentiment classification for big data analysis using naive bayes

classifier. In *2013 IEEE international conference on big data*, pages 99–104. IEEE.

[90] Liu, B. and Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer.

[91] Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., and Watkins, C. (2002). Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444.

[92] Lu, B., Ott, M., Cardie, C., and Tsou, B. K. (2011). Multi-aspect sentiment analysis with topic models. In *2011 IEEE 11th international conference on data mining workshops*, pages 81–88. IEEE.

[93] Lu, Y., Krüger, R., Thom, D., Wang, F., Koch, S., Ertl, T., and Maciejewski, R. (2014). Integrating predictive analytics and social media. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 193–202. IEEE.

[94] Luo, Z., Osborne, M., and Wang, T. (2012). Opinion retrieval in twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 6.

[95] Maipradit, R., Hata, H., and Matsumoto, K. (2019). Sentiment classification using n-gram inverse document frequency and automated machine learning. *IEEE Software*, 36(5):65–70.

[96] Matsumoto, S., Takamura, H., and Okumura, M. (2005). Sentiment classification using word sub-sequences and dependency subtrees. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 301–311. Springer.

[97] Mccallum, A. and Nigam, K. (2001). A comparison of event models for naive bayes text classification. *Work Learn Text Categ*, 752.

[98] Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

[99] Meharia, P. and Agarwal, D. (2018). *Securing the Human Cloud: Applying Biometrics to Wearable Technology*, pages 262–276.

[100] Mishra, N. and Jha, C. (2012). Classification of opinion mining techniques. *International Journal of Computer Applications*, 56(13).

[101] Mishra, P., Pilli, E., Varadharajan, V., and Tupakula, U. (2016). Intrusion detection techniques in cloud environment: A survey. *Journal of Network and Computer Applications*, 77.

[102] Moghaddam, F. F., Yezdanpanah, M., Khodadadi, T., Ahmadi, M., and Eslami, M. (2014). Vdci: Variable data classification index to ensure data protection in cloud computing environments. In *2014 IEEE Conference on Systems, Process and Control (ICSPC 2014)*, pages 53–57. IEEE.

[103] Mortazavi, S., Pour, A., and Kato, T. (2011). An efficient distributed group key management using hierarchical approach with diffie-hellman and symmetric algorithm: Dhsa. pages 49–54.

[104] Muhammad, W. (2016). Implementation of 163-bit elliptic curve diffie hellman key exchange protocol using bigdigits arithmetic. *International Journal of Advanced Trends in Computer Science and Engineering*, 5:65–70.

[105] Myllymäki, P. and Tirri, H. (1993). Bayesian case-based reasoning with neural networks. pages 422 – 427 vol.1.

[106] Nasukawa, T. and Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 70–77.

[107] Neethu, M. and Rajasree, R. (2013). Sentiment analysis in twitter using machine learning techniques. pages 1–5.

[108] Neri, F., Aliprandi, C., Capeci, F., Cuadros, M., and By, T. (2012). Sentiment analysis on social media. *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 919–926.

[109] Nigam, K., Laertyy, J., and McCallumzy, A. (1999). Using maximum entropy for text classication.

[110] Nigam, K., Mccallum, A., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39:103–134.

[111] Pak, A. and Paroubek, P. (2010). Twitter based system: Using twitter for disambiguating sentiment ambiguous adjectives. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 436–439.

[112] Parikh, R. and Movassate, M. (2009). Sentiment analysis of user-generated twitter updates using various classification techniques. *CS224N Final Report*, 118.

[113] Patel, N. and Singh, D. (2015). An algorithm to construct decision tree for machine learning based on similarity factor. *International Journal of Computer Applications*, 111:22–26.

[114] Patodkar, V. and I.R, S. (2016). Twitter as a corpus for sentiment analysis and opinion mining. *IJARCCE*, 5:320–322.

[115] Pazzani, M. and Billsus, D. (1997). Learning and revising user profiles: The identification of interesting web sites. *Machine Learning*, 27:313–331.

[116] Permatasari, R. I., Fauzi, M. A., Adikara, P. P., and Sari, E. D. L. (2018). Twitter sentiment analysis of movie reviews using ensemble features based naïve bayes. In *2018 International Conference on Sustainable Information Engineering and Technology (SIET)*, pages 92–95. IEEE.

[117] Prabhakar, D. M. and Joseph, K. S. (2013). A new approach for providing data security and secure data transfer in cloud computing. *International Journal of Computer Trends and Technology (IJCTT) pp*, pages 1202–120.

[118] Prabhu, J., Sudharshan, M., Mohan, S., and Prasad, G. (2010). Augmenting rapid clustering method for social network analysis. pages 407–408.

[119] Rafaeli, S. and Hutchison, D. (2003). A survey of key management for secure group communication. *ACM Comput. Surv.*, 35:309–329.

[120] Rahnama, A. H. A. (2014). Distributed real-time sentiment analysis for big data social streams. In *2014 International conference on*

control, decision and information technologies (CoDIT), pages 789–794. IEEE.

[121] Rajesh, S. and Michael, J. (2015). Effectiveness of social media in education. *International Journal of Innovative Research in Advanced Engineering*, 10(2):2349–2163.

[122] Rewagad, P. and Pawar, Y. (2013). Use of digital signature with diffie hellman key exchange and aes encryption algorithm to enhance data security in cloud computing. pages 437–439.

[123] Rocchio, J. (1971). *Relevance Feedback in Information Retrieval*, pages 313–323.

[124] Rodrigues Barbosa, G. A., Silva, I. S., Zaki, M., Meira Jr, W., Prates, R. O., and Veloso, A. (2012). Characterizing the effectiveness of twitter hashtags to detect and track online population sentiment. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, pages 2621–2626.

[125] Roy, I., Setty, S., Kilzer, A., Shmatikov, V., and Witchel, E. (2010). Airavat: Security and privacy for mapreduce. pages 297–312.

[126] Ruck, D. W., Rogers, S. K., and Kabrisky, M. (1990). Feature selection using a multilayer perceptron. *Journal of Neural Network Computing*, 2(2):40–48.

[127] Sagar, V., Kumar, K., and Vishnoi, V. (2019). A comparative analysis of cryptographic algorithms. 6:358–373.

[128] Sahay, S. (2021). Support vector machines and document classification.

[129] Saif, H. and Alani, H. (2012). Semantic sentiment analysis of twitter. volume 7649, pages 508–524.

[130] Saleh, M. R., Martín-Valdivia, M. T., Montejo-Ráez, A., and Ureña-López, L. (2011). Experiments with svm to classify opinions in different domains. *Expert Systems with Applications*, 38(12):14799–14804.

[131] Saouli, H., Ghamri, A., Merizig, A., and Kazar, O. (2017). A new cloud computing approach based svm for relevant data extraction.

[132] Sengupta, N. (2013). Designing of cryptography based security system for cloud computing.

[133] Shaikh, R. and Mukundan, S. (2015). Data classification for achieving security in cloud computing. *Procedia Computer Science*, 45:493–498.

[134] Sharif, M. H. and Gursoy, O. (2018). Parallel computing for artificial neural network training using java native socket programming. *Periodicals of Engineering and Natural Sciences*, 6.

[135] Singh, A., Nandal, A., and Malik, S. (2012). Implementation of caesar cipher with rail fence for enhancing data security.

[136] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

[137] Somani, U., Lakhani, K., and Mundra, M. (2010). Implementing digital signature with rsa encryption algorithm to enhance the data security of cloud in cloud computing. pages 211 – 216.

[138] Sood, S. (2012). A combined approach to ensure data security in cloud computing. *Journal of Network and Computer Applications*, 35:1831–1838.

[139] Stergiou, C., Psannis, K., Kim, B.-G., and Gupta, B. B. (2016). Secure integration of iot and cloud computing. *Future Generation Computer Systems*, 78.

[140] Suki, M., Rthi, J., and Vizhi, K. (2015). Decision making using sentiment analysis from twitter. *International Journal of Innovative Research in Computer and Communication Engineering*, 02:7171–7177.

[141] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

[142] Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., and Li, P. (2011). User-level sentiment analysis incorporating social networks.

[143] Tan, S., Cheng, X., Wang, Y., and Xu, H. (2009). Adapting naive bayes to domain adaptation for sentiment analysis. volume 5478, pages 337–349.

[144] Tan, S., Li, Y., Sun, H., Guan, Z., Yan, X., Bu, J., Chen, C., and He, X. (2014). Interpreting the public sentiment variations on twitter. *Knowledge and Data Engineering, IEEE Transactions on*, 26:1158–1170.

[145] Terveen, L., Hill, W., Amento, B., McDonald, D., and Creter, J. (1997). Phoaks: A system for sharing recommendations. *Communications of the ACM*, 40(3):59–62.

[146] Thambiraja, E., Ramesh, G., and Umarani, D. R. (2012). A survey on various most common encryption techniques. *International journal of advanced research in computer science and software engineering*, 2(7).

[147] Tong, S. and Koller, D. (2001). Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66.

[148] Tripathi, G. and Naganna, S. (2015). Feature selection and classification approach for sentiment analysis. *Machine Learning and Applications: An International Journal*, 2(2):1–16.

[149] Turney, P. D. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *arXiv preprint cs/0212032*.

[150] Velez Zea, A., Barrera Ramírez, J. F., and Torroba, R. (2017). Cryptographic salting for security enhancement of double random phase encryption schemes. *Journal of Optics*, 19.

[151] Wan, C. H., Lee, L. H., Rajkumar, R., and Isa, D. (2012). A hybrid text classification approach with low dependency on parameter by integrating k-nearest neighbor and support vector machine. *Expert Systems with Applications*, 39(15):11880–11888.

[152] Wang, C., Wang, Q., Ren, K., and Lou, W. (2009). Ensuring

data storage security in cloud computing. *IACR Cryptology ePrint Archive*, 2009:81.

[153] Wang, C., Xiao, Z., Liu, Y., Xu, Y., Zhou, A., and Zhang, K. (2013). Sentiview: Sentiment analysis and visualization for internet popular topics. *IEEE transactions on human-machine systems*, 43(6):620–630.

[154] Wang, H., Lu, Y., and Zhai, C. (2011). Latent aspect rating analysis without aspect keyword supervision. pages 618–626.

[155] Wang, Z., Joo, V., Tong, C., and Chan, D. (2014). Issues of social data analytics with a new method for sentiment analysis of social media data. In *2014 IEEE 6th International Conference on Cloud Computing Technology and Science*, pages 899–904. IEEE.

[156] Wiebe, J. (1990). Identifying subjective characters in narrative. In *COLNG 1990 Volume 2: Papers presented to the 13th International Conference on Computational Linguistics.*

[157] Wilson, T., Wiebe, J., and Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics*, 35(3):399–433.

[158] Xuan, J., He, J., Ren, Z., Yan, J., and Luo, Z. (2017). Automatic bug triage using semi-supervised text classification.

[159] Zardari, M. A., Jung, L. T., and Zakaria, N. (2014). K-nn classifier for data confidentiality in cloud computing. In *2014 International Conference on Computer and Information Sciences (ICCOINS)*, pages 1–6. ieee.

[160] Zhang, X., Zhao, J., and Lecun, Y. (2015). Character-level convolutional networks for text classification.

LIST OF PUBLICATIONS

- Gitanjali , Kamlesh Lakhwani, “Securing Big Data over Cloud Using Classification and Encryption Techniques”, IJRECE VOL. 6 ISSUE 2 APR-JUNE 2018.[Published]
- Gitanjali , Kamlesh Lakhwani, “A Novel Approach of Sensitive Data Classification using Convolution Neural Network and Logistic Regression Techniques” in Scopus based International Journal of Innovative Technology and Exploring Engineering (IJITEE) VOL. 8 ISSUE 8, JUNE 2019.(Scopus Indexed) [Published].
- Gitanjali Gupta , Kamlesh Lakhwani (2020) ’Improved Encryption of Big Data by Shift Homomorphic with ECDH Approach’, Test Engineering and Management Journal, 83 (March - April 2020), pp. 25416- 25424. (Scopus Indexed) [Published]
- Gitanjali Gupta , Kamlesh Lakhwani (2020) ’BIG DATA CLASSIFICATION TECHNIQUES: A SYSTEMATIC LITERATURE REVIEW’, Journal of Natural Remedies, 21 No. 2(Special Issue-1 : Recent advancement in the Science, Engineering and Technology Section). (Scopus Indexed) [Published].
- Gitanjali, Kamlesh Lakhwani. "An enhanced intelligent classification approach to improve the encryption of big data". International Conference on Artificial Intelligence and Machine learning (ICAIML 2020), 2020, IOP Conf. Ser.: Mater. Sci. Eng. 1049 012008. (IEEE) [Published].

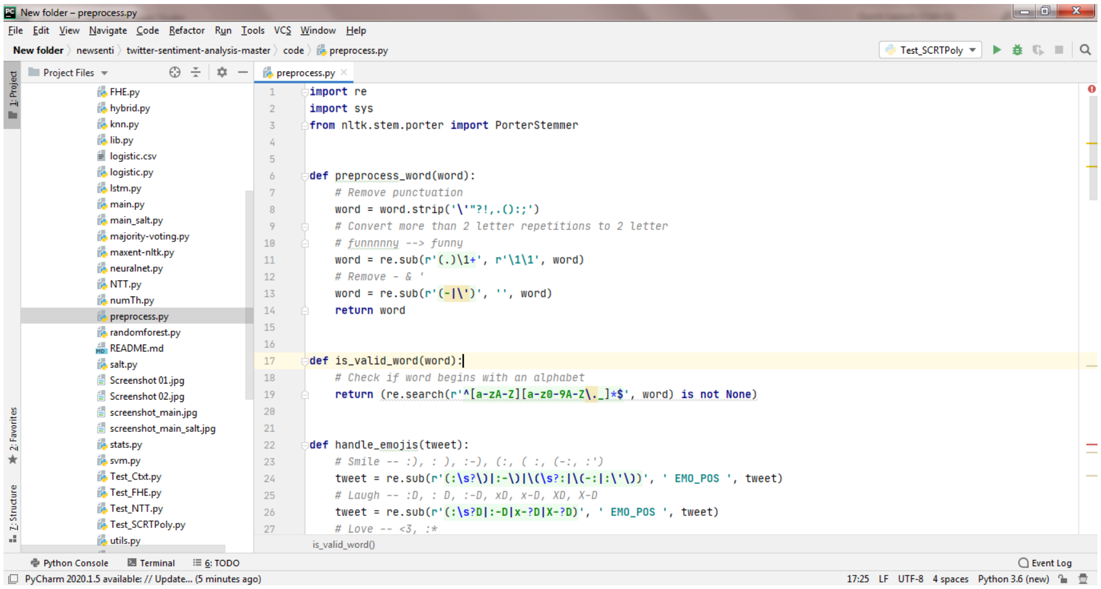
- Gitanjali , Kamlesh Lakhwani, A novel Security Aware Sensitive Encrypted Storage approach to improve the encryption of big data, Journal of Multimedia Tools and applications(SCI) [Communicated].

APPENDIX A

The Appendix Section is the step by step explanation of the entire research process. This section is basically used to show the proposed framework which includes the snapshots of the following steps:-

- Data Preprocessing
- Features Extraction
- Accuracy rate calculation of various Classifiers
- Encyption Keys Generation
- Implementation of Encryption Technique
- Implementation of Decryption Technique

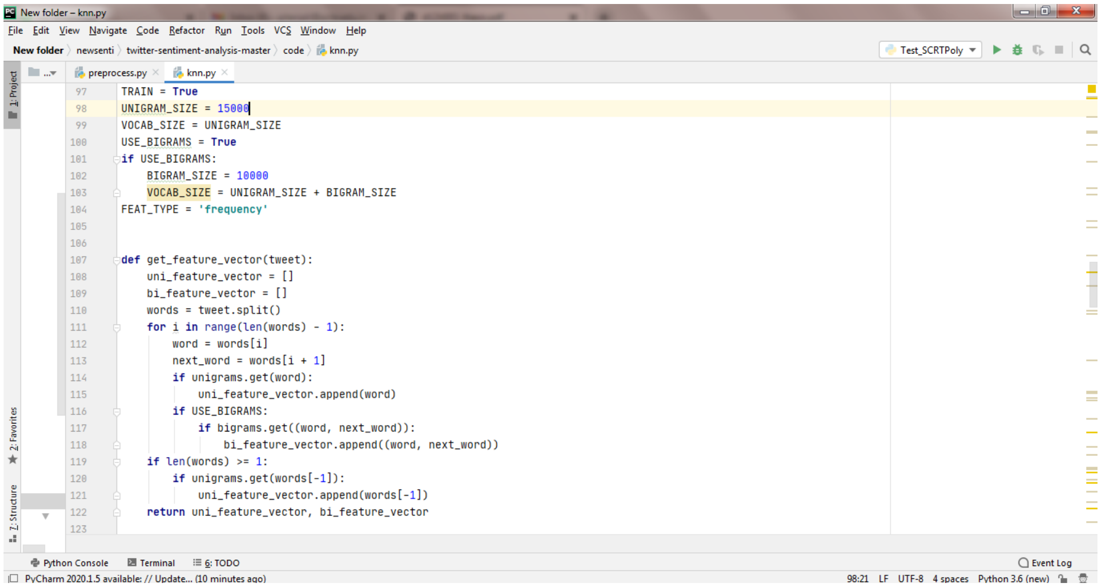
STEP-1: Data Preprocessing



```python
import re
import sys
from nltk.stem.porter import PorterStemmer


def preprocess_word(word):
    # Remove punctuation
    word = word.strip('\'"?!,.():;')
    # Convert more than 2 letter repetitions to 2 letter
    # funnnnny --> funny
    word = re.sub(r'(.)\1+', r'\1\1', word)
    # Remove - & '
    word = re.sub(r'(-|\')', '', word)
    return word


def is_valid_word(word):
    # Check if word begins with an alphabet
    return (re.search(r'^[a-zA-Z][a-z0-9A-Z\._]*$', word) is not None)


def handle_emojis(tweet):
    # Smile -- :), : ), :-), (:, ( :, (-:, :')
    tweet = re.sub(r'(:\s?\)|:-\)|\(\s?:|\(-:|:\'\))', ' EMO_POS ', tweet)
    # Laugh -- :D, : D, :-D, xD, x-D, XD, X-D
    tweet = re.sub(r'(:\s?D|:-D|x-?D|X-?D)', ' EMO_POS ', tweet)
    # Love -- <3, :*
```

Figure A. 1: Data preprocessing

STEP-2: Features Extraction using N-GRAM

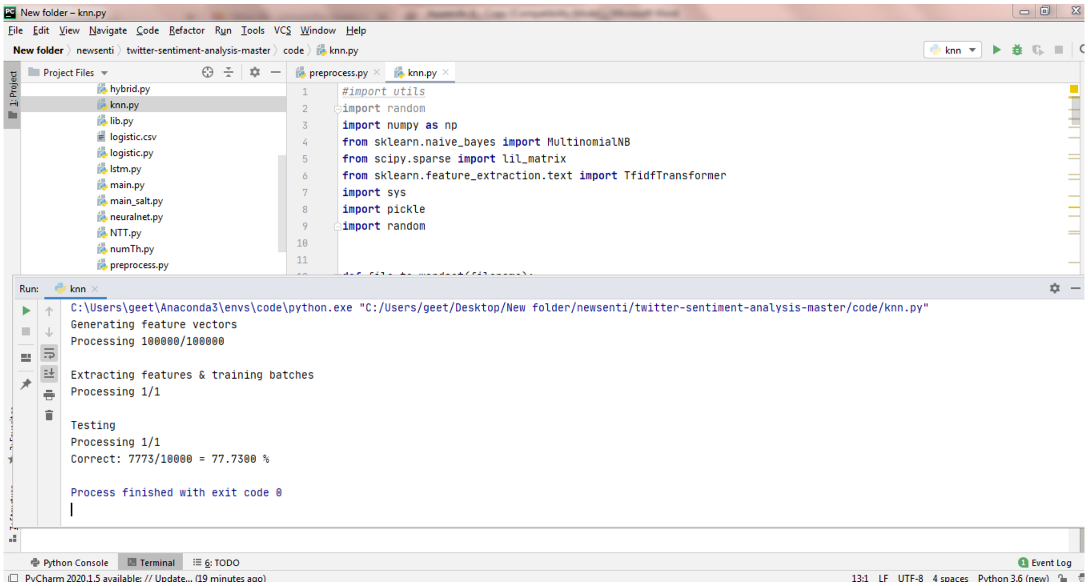

New folder – knn.py
File Edit View Navigate Code Refactor Run Tools VCS Window Help
New folder 〉 newsenti 〉 twitter-sentiment-analysis-master 〉 code 〉 knn.py
Test_SCRTPoly
preprocess.py
knn.py
Python Console
Terminal
6: TODO
Event Log
PyCharm 2020.1.5 available: // Update... (10 minutes ago)
98:21 LF UTF-8 4 spaces Python 3.6 (new)

```python
TRAIN = True
UNIGRAM_SIZE = 15000
VOCAB_SIZE = UNIGRAM_SIZE
USE_BIGRAMS = True
if USE_BIGRAMS:
    BIGRAM_SIZE = 10000
    VOCAB_SIZE = UNIGRAM_SIZE + BIGRAM_SIZE
FEAT_TYPE = 'frequency'


def get_feature_vector(tweet):
    uni_feature_vector = []
    bi_feature_vector = []
    words = tweet.split()
    for i in range(len(words) - 1):
        word = words[i]
        next_word = words[i + 1]
        if unigrams.get(word):
            uni_feature_vector.append(word)
        if USE_BIGRAMS:
            if bigrams.get((word, next_word)):
                bi_feature_vector.append((word, next_word))
    if len(words) >= 1:
        if unigrams.get(words[-1]):
            uni_feature_vector.append(words[-1])
    return uni_feature_vector, bi_feature_vector
```

Figure A. 2: Features extraction

STEP-3: Accuracy rate of various classifiers

3A) Accuracy rate with KNN

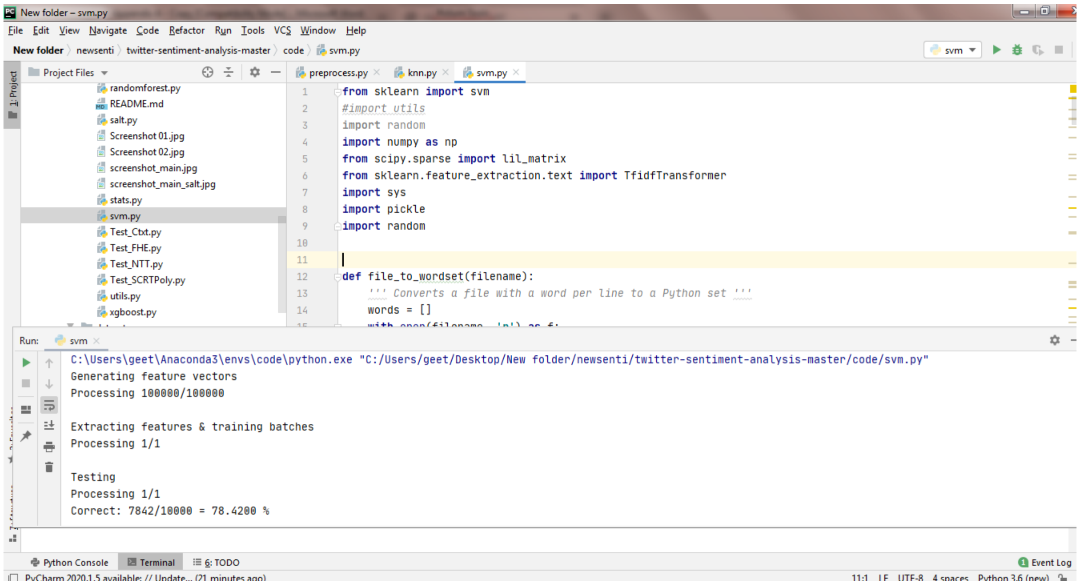

New folder – knn.py
File Edit View Navigate Code Refactor Run Tools VCS Window Help
New folder 〉 newsenti 〉 twitter-sentiment-analysis-master 〉 code 〉 knn.py
knn
Project Files
hybrid.py
knn.py
lib.py
logistic.csv
logistic.py
lstm.py
main.py
main_salt.py
neuralnet.py
NTT.py
numTh.py
preprocess.py
preprocess.py
knn.py
Run: knn
Python Console
Terminal
6: TODO
Event Log
PyCharm 2020.1.5 available: // Update... (19 minutes ago)
13:1 LF UTF-8 4 spaces Python 3.6 (new)

```
#import utils
import random
import numpy as np
from sklearn.naive_bayes import MultinomialNB
from scipy.sparse import lil_matrix
from sklearn.feature_extraction.text import TfidfTransformer
import sys
import pickle
import random
```

```
C:\Users\geet\Anaconda3\envs\code\python.exe "C:/Users/geet/Desktop/New folder/newsenti/twitter-sentiment-analysis-master/code/knn.py"
Generating feature vectors
Processing 100000/100000

Extracting features & training batches
Processing 1/1

Testing
Processing 1/1
Correct: 7773/10000 = 77.7300 %

Process finished with exit code 0
```

Figure A. 3: Accuracy rate with KNN

3B) Accuracy rate with SVM

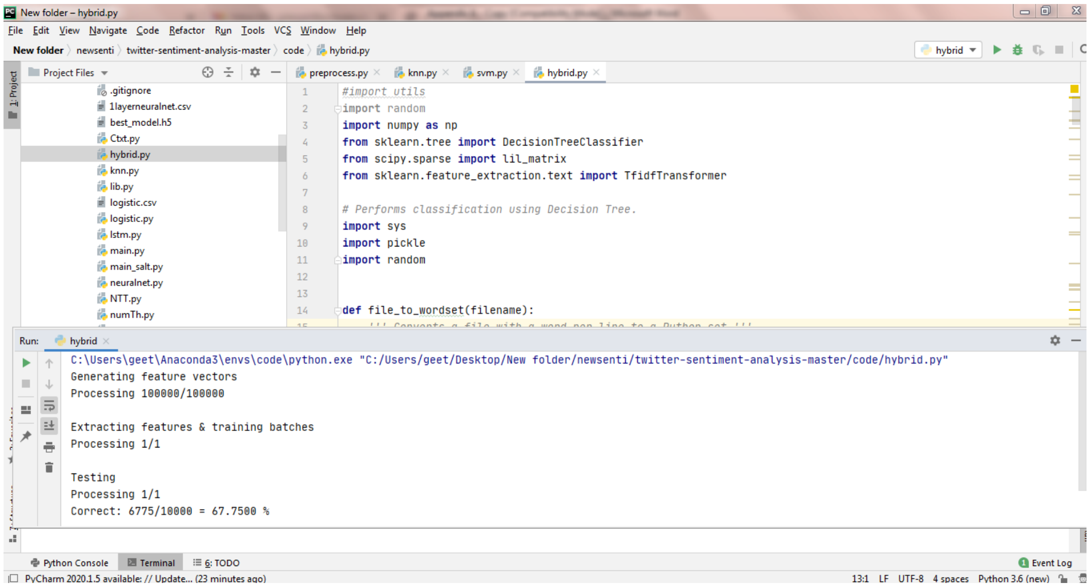


```
from sklearn import svm
#import utils
import random
import numpy as np
from scipy.sparse import lil_matrix
from sklearn.feature_extraction.text import TfidfTransformer
import sys
import pickle
import random


def file_to_wordset(filename):
    ''' Converts a file with a word per line to a Python set '''
    words = []
```

```
C:\Users\geet\Anaconda3\envs\code\python.exe "C:/Users/geet/Desktop/New folder/newsenti/twitter-sentiment-analysis-master/code/svm.py"
Generating feature vectors
Processing 100000/100000

Extracting features & training batches
Processing 1/1

Testing
Processing 1/1
Correct: 7842/10000 = 78.4200 %
```

Figure A. 4: Accuracy rate with SVM

3C) Accuracy rate with Decision Tree

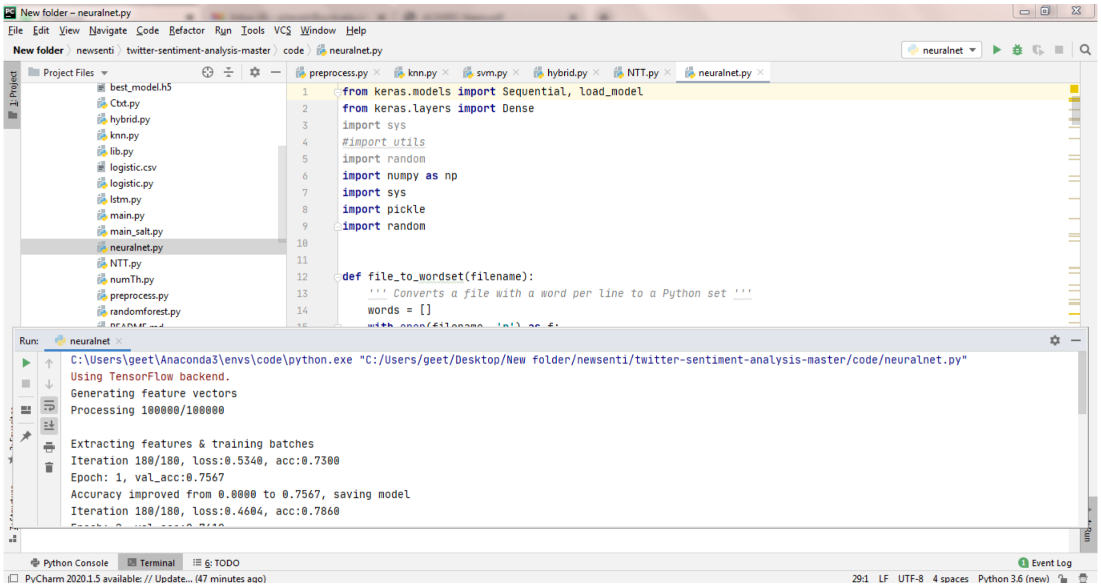


```python
#import utils
import random
import numpy as np
from sklearn.tree import DecisionTreeClassifier
from scipy.sparse import lil_matrix
from sklearn.feature_extraction.text import TfidfTransformer

# Performs classification using Decision Tree.
import sys
import pickle
import random


def file_to_wordset(filename):
    ''' Converts a file with a word per line to a Python set '''
```

```
C:\Users\geet\Anaconda3\envs\code\python.exe "C:/Users/geet/Desktop/New folder/newsenti/twitter-sentiment-analysis-master/code/hybrid.py"
Generating feature vectors
Processing 100000/100000

Extracting features & training batches
Processing 1/1

Testing
Processing 1/1
Correct: 6775/10000 = 67.7500 %
```

Figure A. 5: Accuracy rate with Decision Tree

3D) Accuracy rate with Neural Network

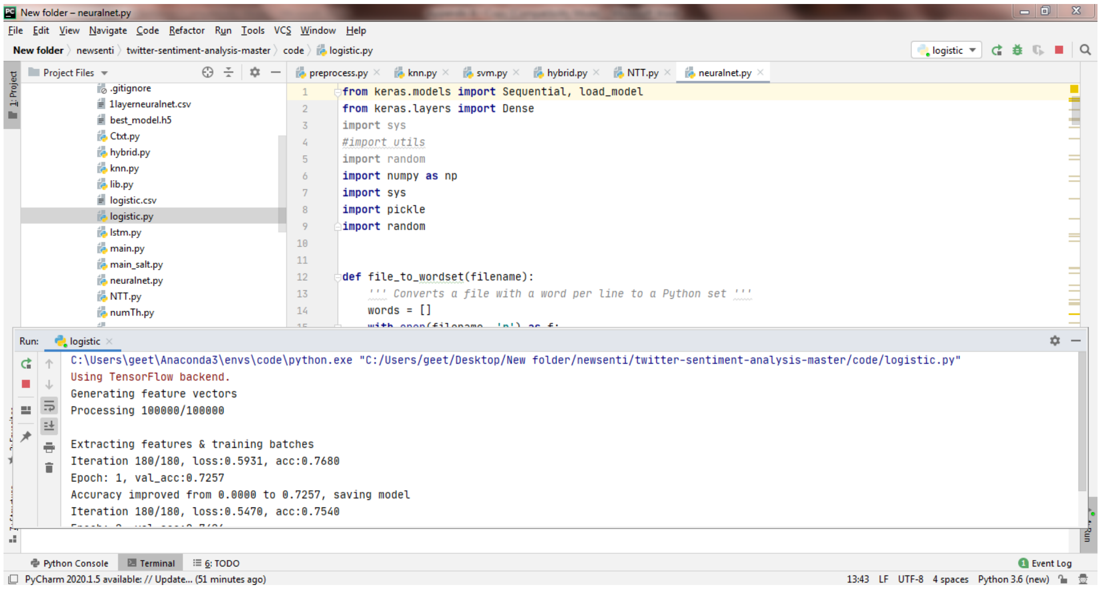


```python
from keras.models import Sequential, load_model
from keras.layers import Dense
import sys
#import utils
import random
import numpy as np
import sys
import pickle
import random


def file_to_wordset(filename):
    ''' Converts a file with a word per line to a Python set '''
    words = []
    with open(filename, 'r') as f:
```

```
C:\Users\geet\Anaconda3\envs\code\python.exe "C:/Users/geet/Desktop/New folder/newsenti/twitter-sentiment-analysis-master/code/neuralnet.py"
Using TensorFlow backend.
Generating feature vectors
Processing 100000/100000

Extracting features & training batches
Iteration 180/180, loss:0.5340, acc:0.7300
Epoch: 1, val_acc:0.7567
Accuracy improved from 0.0000 to 0.7567, saving model
Iteration 180/180, loss:0.4604, acc:0.7860
```

Figure A. 6: Accuracy rate with Neural Network

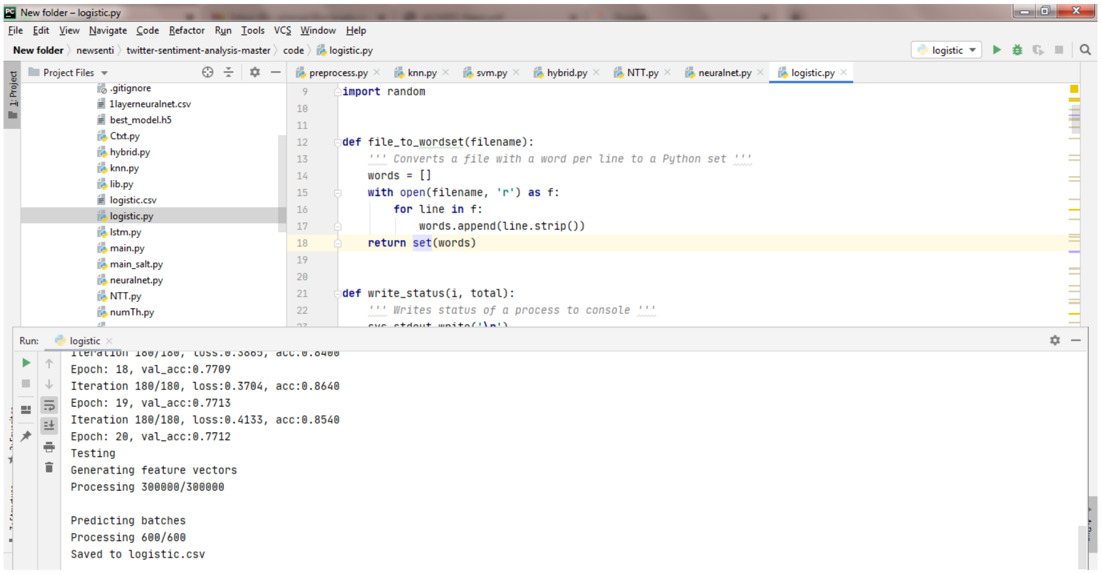
3E) Accuracy rate with CNN-LR


New folder – neuralnet.py
File Edit View Navigate Code Refactor Run Tools VCS Window Help
New folder 〉 newsenti 〉 twitter-sentiment-analysis-master 〉 code 〉 logistic.py
logistic
Project Files
.gitignore
1layerneuralnet.csv
best_model.h5
Ctxt.py
hybrid.py
knn.py
lib.py
logistic.csv
logistic.py
lstm.py
main.py
main_salt.py
neuralnet.py
NTT.py
numTh.py
preprocess.py
svm.py
Run: logistic
Python Console
Terminal
6: TODO
Event Log
PyCharm 2020.1.5 available: // Update... (51 minutes ago)
13:43 LF UTF-8 4 spaces Python 3.6 (new)

```python
from keras.models import Sequential, load_model
from keras.layers import Dense
import sys
#import utils
import random
import numpy as np
import sys
import pickle
import random


def file_to_wordset(filename):
    ''' Converts a file with a word per line to a Python set '''
    words = []
    with open(filename, 'r') as f:
```

```
C:\Users\geet\Anaconda3\envs\code\python.exe "C:/Users/geet/Desktop/New folder/newsenti/twitter-sentiment-analysis-master/code/logistic.py"
Using TensorFlow backend.
Generating feature vectors
Processing 100000/100000

Extracting features & training batches
Iteration 180/180, loss:0.5931, acc:0.7680
Epoch: 1, val_acc:0.7257
Accuracy improved from 0.0000 to 0.7257, saving model
Iteration 180/180, loss:0.5470, acc:0.7540
```

Figure A. 7: Accuracy rate with proposed CNN-Logistic Regression



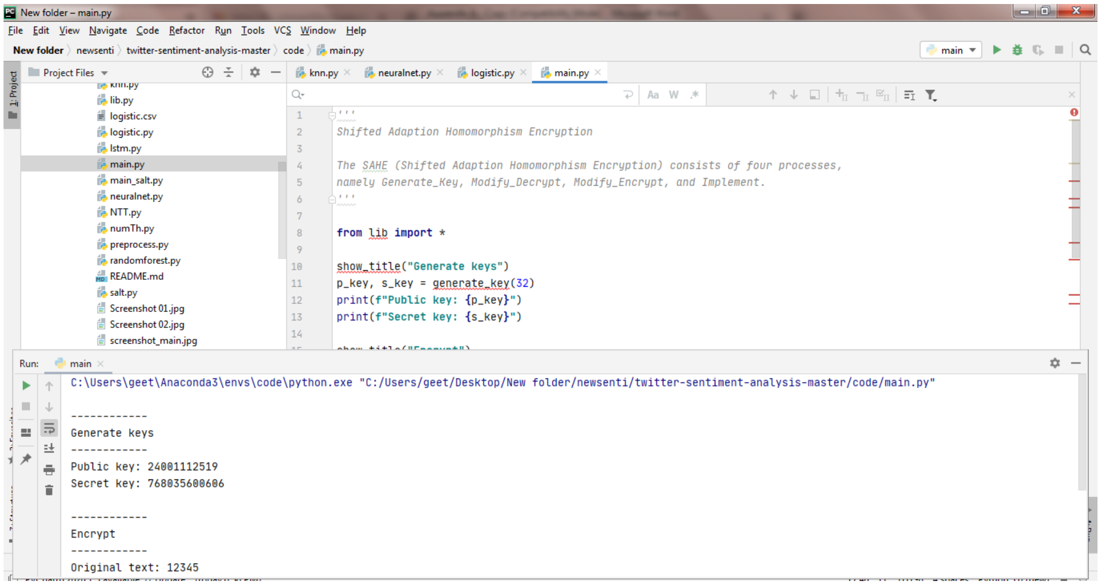
```python
import random


def file_to_wordset(filename):
    ''' Converts a file with a word per line to a Python set '''
    words = []
    with open(filename, 'r') as f:
        for line in f:
            words.append(line.strip())
    return set(words)


def write_status(i, total):
    ''' Writes status of a process to console '''
    sys.stdout.write('\n')
```

```
Iteration 180/180, loss:0.3885, acc:0.8400
Epoch: 18, val_acc:0.7709
Iteration 180/180, loss:0.3704, acc:0.8640
Epoch: 19, val_acc:0.7713
Iteration 180/180, loss:0.4133, acc:0.8540
Epoch: 20, val_acc:0.7712
Testing
Generating feature vectors
Processing 300000/300000

Predicting batches
Processing 600/600
Saved to logistic.csv
```

Figure A. 8: Increased Epochs of proposed CNN-Logistic Regression

STEP-4: Generation of Encryption Keys

In this step public key and secret key will be generated.

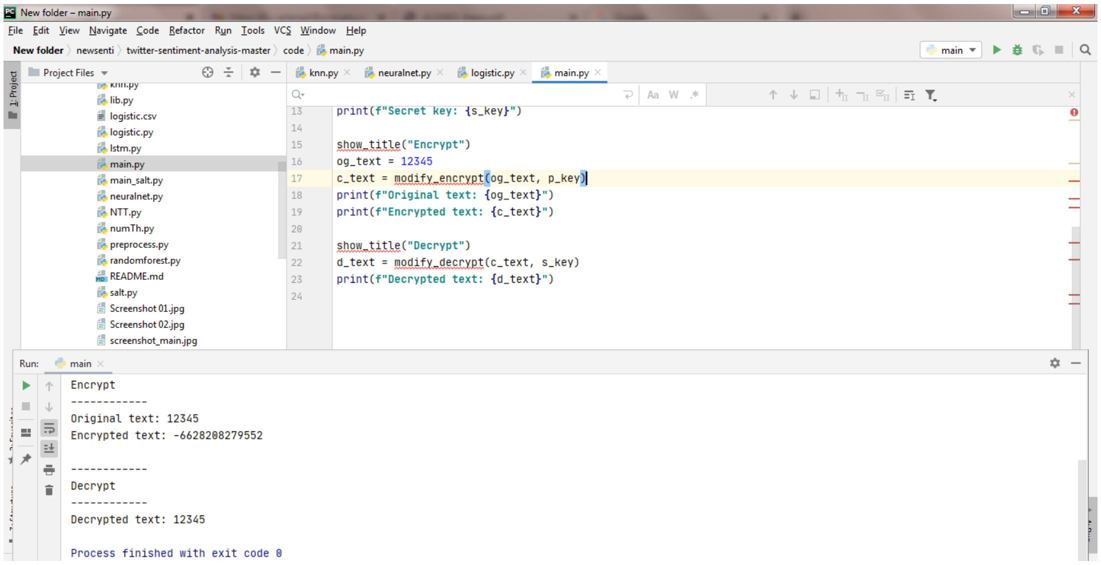


```
'''
Shifted Adaption Homomorphism Encryption

The SAHE (Shifted Adaption Homomorphism Encryption) consists of four processes,
namely Generate_Key, Modify_Decrypt, Modify_Encrypt, and Implement.
'''

from lib import *

show_title("Generate keys")
p_key, s_key = generate_key(32)
print(f"Public key: {p_key}")
print(f"Secret key: {s_key}")

```

```
C:\Users\geet\Anaconda3\envs\code\python.exe "C:/Users/geet/Desktop/New folder/newsenti/twitter-sentiment-analysis-master/code/main.py"

------------
Generate keys
------------
Public key: 24001112519
Secret key: 768035600606

------------
Encrypt
------------
Original text: 12345
```

Figure A. 9: Generating Public and Secret Key

STEP-5: Implementation of SAHE Encryption



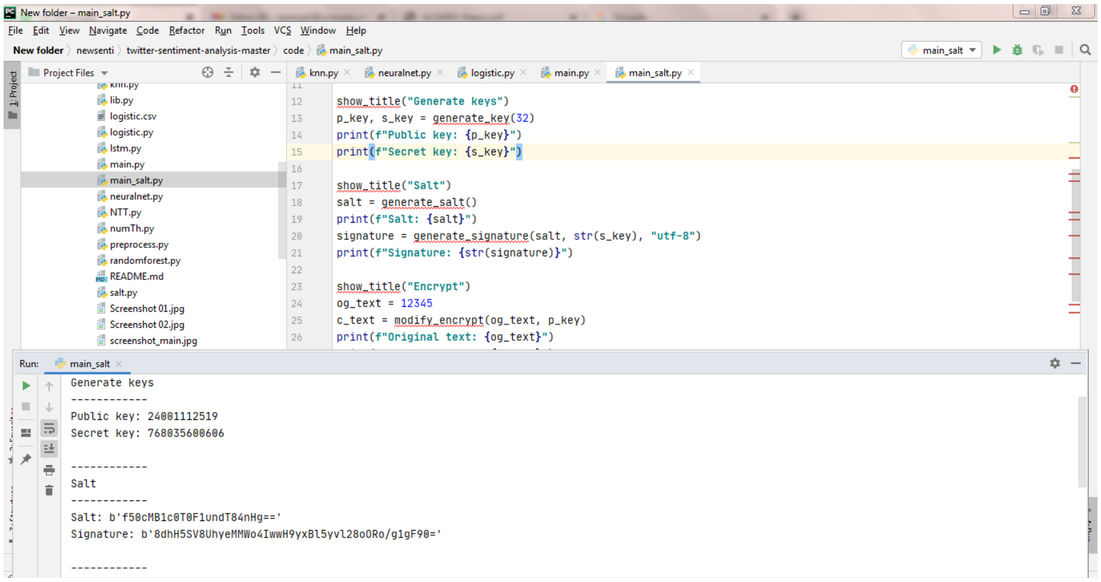
```
print(f"Secret key: {s_key}")

show_title("Encrypt")
og_text = 12345
c_text = modify_encrypt(og_text, p_key)
print(f"Original text: {og_text}")
print(f"Encrypted text: {c_text}")

show_title("Decrypt")
d_text = modify_decrypt(c_text, s_key)
print(f"Decrypted text: {d_text}")

```

```
Encrypt
------------
Original text: 12345
Encrypted text: -6628208279552

------------
Decrypt
------------
Decrypted text: 12345

Process finished with exit code 0
```

Figure A. 10: Shifted Adoption Homomorphism Encryption Implementation

STEP-6: Key Generation with Salting Technique

In this step key will be generated using salting technique.



```
show_title("Generate keys")
p_key, s_key = generate_key(32)
print(f"Public key: {p_key}")
print(f"Secret key: {s_key}")

show_title("Salt")
salt = generate_salt()
print(f"Salt: {salt}")
signature = generate_signature(salt, str(s_key), "utf-8")
print(f"Signature: {str(signature)}")

show_title("Encrypt")
og_text = 12345
c_text = modify_encrypt(og_text, p_key)
print(f"Original text: {og_text}")
```

```
Generate keys
------------
Public key: 24001112519
Secret key: 768035600606

------------
Salt
------------
Salt: b'f50cMB1c0T0F1undT84nHg=='
Signature: b'8dhH5SV8UhyeMMWo4IwwH9yxBl5yvl28oORo/g1gF90='

------------
```

Figure A. 11: Key Generation with Salting Technique