JADAVPUR UNIVERSITY

MASTER DEGREE THESIS

Studies for Detecting of Fake & Real News

A thesis submitted in partial fulfillment of the requirements
For the degree of Master of Technology in Computer Technology

By
PRABIR BERA
University Roll Number: 001810504023
Examination Roll Number: M6TCT22001
Registration Number: 145302 of 2018-19

Under the Guidance of
Prof. Diganta Saha
Department of Computer Science and Engineering
Faculty of Engineering and Technology
Jadavpur University
Kolkata-700032

Aug, 2022

Declaration of Authorship

I, Prabir Bera, declare that this thesis titled, “Studies for Detecting of Fake & Real News” and the works presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a master’s degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly at- tribute.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

To whom it may concern

This is to certify that Prabir Bera has satisfactorily completed the work in this thesis entitled “Studies for Detecting of Fake & Real News”, University Roll Number: 001810504023, Examination Roll Number: M6TCT22001, Registration Number: 145302 of 2018-2019. It is a bona fide piece of work carried out under my super vision at Jadavpur University, Kolkata-700032,for partial fulfillment of the requirements for the degree of Master of Technology in Computer Technology under the Department of Computer Science & Engineering, Jadavpur University for the academic session2018-2022.

Dr. Diganta Saha
Professor
Department of Computer Science
& Engineering
Jadavpur University
Kolkata- 700032.

HOD
Department of Computer Science
& Engineering
Jadavpur University
Kolkata- 700032.

DEAN Faculty of Engineering &Technology
Jadavpur University
Kolkata- 700032.

Certificate of Approval

(Only in case the thesis is approved)

This is to certify that the thesis entitled “Studies for Detecting of Fake & Real News” is a bona fide record of work carried out by Prabir Bera, University Roll Number: 001810504023, Examination Roll Number: : M6TCT22001, Registration Number: 145302 of 2018-2019, in partial fulfillment of the requirements for the award of the degree of Master of Technology in Computer Technology under Department of Computer Science & Engineering, Jadavpur University for the academic session 2018-2022.It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn there in but approve the thesis only for the purpose for which it has been submitted.

(Signature of the external Examiner)

Date:

(Signature of the Examiner)

Date:

Jadavpur University

Abstract

Master of Technology in Computer Technology in

Studies for Detecting of Fake & Real News

By
Prabir Bera
University Roll Number: 001810504023
Examination Roll Number: M6TCT22001
Registration Number: 145302 of 2018-2019

Fake information has been spreading in extra numbers and has generated increasingly more misinformation, one of the clearest examples being the USA presidential elections of 2016, for which lots of fake facts become circulated earlier than the votes that progressed the photograph of Donald Trump overs Hilary's Clinton. Because faux information is just too much, it will become essential to apply computational gear to locate them; that is why using algorithms of Machine Learning like undefined a Naive Bayes Model and herbal language processing for the identity of fake information in public facts units is proposed. Spreading fake news over popular social media like Facebook, Twitter, Whats App, etc. creates serious social problems. It is observed that the administration had to stop internet service in a large area to control the spreading of such fake news. Bar in the internet service creates another serious problem for the citizens as we are heavily dependent on the internet nowadays. To detect fake news, this work suggests a machine learning method using the Naïve Bayes classifier. The system can be plugged in with any social media, as it predicts the probability of news to be fake the same can be stopped from spreading farther with the software control in the social media. This model was trained and tested in 3 datasets (2 of the English language and 1 of Bengali language). We have successfully achieved a high accuracy which is 81% for the English language and

93% for the Bengali language. This is an acceptable result comparing the same with the recent works. Using Deep Learning or a combination of classifiers the result may be improved further.

With the increasing usage of wearable smart devices for caption monitoring, providing caption facilities in remote areas are being developed. For an organization managing all these data from Caption devices and other news information becomes difficult in a local database. The situation is similar for large News with lots of departments as well. The cloud computing technologies can provide a low-cost and scalable solution to this huge volume of data. However, with the advantages of a cloud database, come various security issues. Firstly, secure data transfer through a net- work is needed so that data cannot be stolen or tampered. Also data stored in the cloud database needs to be protected because if the cloud provider is entrusted or is attacked by an intruder, data confidentiality and integrity can be lost. Encrypting all the data stored in the cloud database provides the required level of security. However It produces several challenges in searching an encrypted data base. This thesis describes the challenges and presents an approach to secure a Cassandra database used in an entrusted cloud environment. Specifically, the proposed model provides a secure inter face to the data that is stored in a distributed data base in cloud, by tackling threats of both internal and external attackers. It stores news information in encrypted form in the database and also modification-sensitive information is hashed and stored in block chain. Thus, it provides seamless access to the encrypted data for permitted users while limiting the access for other users in the system, internal and external attackers; along with preventing unauthorized data modification. Storage of hashes allows the system to validate any information at any point of time, while access to the hash values is provided with a secure protocol similar to a block chain.

Acknowledgements

On the submission of

Studies for Detecting of Fake & Real News

I wish to express gratitude to the Department of Computer Science & Engineering for sanctioning a thesis work, from Jadavpur University under which this work has been completed.

I would like to convey my sincere gratitude to Dr. Diganta Saha, Professor, Department of Computer Science & Engineering, Jadavpur University for his valuable suggestions throughout the duration of the thesis work. I am really grateful to him for his constant support which helped me to fully involve myself in this work and develop new approaches in the field of Design of Fake News.

Lastly I would like to thanks to all my family members, teachers, classmates, guardians and well-wishers for encouraging and co-operating me throughout the development of this thesis.

I would like to especially thank my parents whose blessings helped me to carry out my thesis in a dedicated way.

Regards,
PRABIR BERA
University Roll Number: 001810504023
Examination Roll Number: M6TCT22001
Registration Number: 145302 of 2018-2019
Department of Computer Science & Engineering,

Jadavpur University

Signed: ____

Date ____

Contents

List of Figures

List of Tables

Chapter 1

Introduction

What is Fake News? -‘Fake News’ is a term that can, in a broader sense, classify news which are either false or click-bait or hate news. It is a kind of news consisting of disinformation and can spread via traditional newsprint media or digital media. With the advent of social networking sites, common people have become more susceptible to this kind of news and knowingly or unknowingly have become an agent to spread this news. Fake news is avoidable as it can refer to a diverse range of disinformation covering topics such as health, environment and economics across all platforms and genres. However, the term is new. Fake news started to roll out since 2016 and it named Collins dictionary official word of the year for 2017. Click bait, misleading headlines, sloppy journalism, biased news targeted towards a person or a community are the causes of creating fake news around people.

Nowadays social media is playing an important role in spreading news whether it is real or fake and people are ready to imbibe it. News in social media uses interesting text, images and videos which attract the readers even more. It is cheap to access and circulated quickly. News through social media has got more popularity than the traditional news of the newspaper. So news reports made with purposely false information are easy to spread. It is considered harmful for the news ecosystem and changes people’s mind towards real news. In a nutshell, fake news and create a negative effect on individuals and as well as in society. To alleviate the effect of this, the detection of fake news now becomes an alluring topic to research.

For the current research three datasets have been used. Two of them were in English where one dataset called Liar has only small statements and the other dataset called Fake News Corpus has headlines as well as body parts. The Bengali dataset which is called Ban Fake News has headlines and body parts too.

Supervised Machine Learning methods have been used on the datasets. The objective of this work is to recognize fake or false news by considering different attributes of the news corpus in the classification process.

This work is one of the very first works on Fake News Detection in the Bengali Language. As this work achieved an accuracy of 93%, we can say it's a very good first step towards the future.

Digitization of essential services paved the way to overcome lots of real-life problems across many sectors such as education, banking, transportation etc. Interest in data-driven banking facilities has been growing rapidly. This resulted in generation huge amount of banking related data across many different banking services and organizations. Moreover different organization related to the healthcare services can contribute to as hared pool of data that is simultaneously generated and accessed by organizations such as education, news and government bodies. This requires efficient and scalable storage and pervasive access to the data. Cloud computing is a natural candidate for storing and processing this huge amount of data. Typically such data is stored in a relational database (RDBMS) such as Oracle, SQL Server, DB2, SQLite, etc. Recent trends show the use of No SQL data bases such as Cassandra, Mongo, H Base, etc. to process such unstructured data that allows more flexibility in data generation and processing.

1.1 Overview

This thesis presents an efficient and secure system model for managing news data stored in a public cloud environment. In short, this model provides a secure interface to the data that is stored in a distributed database, by tackling threats of both internal and external attackers, also provide efficient query management scheme. This model uses encryption to store data securely and thereby protecting its confidentiality. A secure interface is provided to the news for accessing the required data on-demand for both read and write requests. Specifically, the model allows encrypted data to be shared by several concerned users in a secure manner; i.e. it provides seamless access to the encrypted data for permitted users, while limiting the access for other users in the system along with the internal or external attackers.

To preserve data integrity from external attackers and malicious internal users, we propose a system that uses a block chain to check and prevent unauthorized data modification. A block chain is a list of records, called *blocks*; such that each block contains a cryptographic hash of the previous block, a timestamp and some arbitrary data. We also propose an efficient strategy to sear chin the block chain that is applicable to fake news data base systems.

Specifically, hash values of each sensitive data record are stored in a block chain. These hash values of the data cannot be modified due to the immutable property of a block chain. This means, once hash values of data records has been written to a block chain, not even an authorized user can change it. Also, because of the immutable property of a block chain, only the parts of the data that should not be modified are stored in the block chain.

In the data context, there is much information that need not be changed after inserting in the database. We describe a system that handles the security and integrity of such sensitive data, reducing the risk of information leakage.

1.2 Outline

The rest of the thesis is organized as follows:

- Chapter 2: Describes a brief review of literature on security issue cloud database systems.
- Chapter3: Provides a brief introduction of the concepts and the technologies used in this thesis.
- Chapter 3: Describes the proposed system and highlights the specific contributions.
- Chapter 4: Describes the implementation details and reports the performance of the implemented system following the proposed approach.
- Chapter5: Draws the summary of proposed system and shows the directions for future work.

Chapter 2

Literature Survey

There are many approaches data security in public cloud. Kantarciogˇlu and Clifton, 2005 proposed a theoretical overview of a secure database server that provides probabilistic security guarantees. The authors loosely elaborated in the paper about how research in this area should proceed. They presented an efficient encrypted database and query processing model. Some security norms are also proposed in the encrypted database, such as any two tables with the same schema and the same number of tupelos must have indistinguishable encryption methods. Antipathy et al.,2011presented a distributed architecture that provides both privacy as well as fault tolerance to the client. In this paper a query partitioning approach is taken for the query at the client side to the servers, which has satisfies the privacy constraints even after local changes to a partition. However, these most of the proposed approaches in the literature focuses on general data and their applicability is limited for data storage. The limitations are in both in terms of the unique challenges that the properties of the fake data possess, and also in terms of the scope of exploiting domain specific information to make the system more efficient. This thesis focuses on explicitly tackling the problem of querying encrypted data. Also, most of the previous approaches to secure a cloud data base overlook the threat of losing data integrity by the internal attacks, i.e. the possibility of an authorized, but malicious user modifying the stored data. In this chapter overview of the related works in these two directions are provided.

The authors have proposed a method by using HAN to detect fake news in [1]. This model has been evaluated in Brazilian and

Portuguese fake news corpus which is named as Fake.Br. Removing or keeping stop words and varying the word embedding's size showed slightly less accuracy. The authors have proposed a modern approach to detect fake news which achieves high accuracy in [2]. To group news articles in clusters, it has a topic-based classifying mechanism and to extract events it has event extraction mechanism. It compares their events with legitimate events to find the credibility of news. H. Karimi, P. Roy, S. Saba Sadiya , & J. Tang have introduced a framework called MMFD which combines extracted features, multi-source fusion and degrees of the fakeness of articles into a single understandable model in [3] and in [4] the authors have proposed a new framework named EANN. This framework is made of three elements, a feature extractor which extracts textual and visual features from posts, a detector which learns the differentiable representations and event discriminator to distinguish between fake and not fake news. In [5] the authors have proposed a network which is named as TI-CNN. The method projects the latent and explicit features to a single space. This model can be trained and tested with both the image and text data at the same time. H. Ahmed, I. Traore, & S. Saad have proposed a method which uses an analysis technique called N-Gram with some techniques of Artificial Intelligence in [6]. After experimenting with many of feature extraction techniques and classification techniques they have found that TF-IDF feature extraction technique and LSVM classification technique gives the best performance. The authors have used an automated detector which is based on the Artificial Intelligence technique deep learning and a 3HAN to achieve speed and accuracy in [7]. The network consists of 3 tiers, one tier for words, one tier for sentences made of those words and one tier for news vector. This method provides an easy to understand outcome using the values given to components of the used article.

In [8] the authors have proposed a method called CSI which consists of 3 models, for Capturing, for Scoring and for Integrating. The first module uses RNN to capture the user activity, the second module learns the characteristics of the source. Then by integrating both the modules with the Integrate module the model can classify whether the article is not fake or fake. W. Y. Wang. Has proposed a benchmark dataset consists of manually labelled and decade-long short statements of various Contexts, ‘LIAR’ in [9]. The dataset yielded an improved result of the deep learning model. In [10] M. Granik, and V. Mesyura have proposed an elementary approach for fake news detection using the Naive Bayes technique. The software system was tested against a dataset created using Facebook posts and achieved a decent result. The authors have introduced a model that uses a hybrid approach which is a combination of ML techniques and linguistic cue with behavioral data in [11].

2.1 Proposed work

We have tried and tested many classification techniques like Logistic Regression, K-Nearest Neighbors, Naive Bayes and many more. Naive Bayes classifier stands out in the list in terms of accuracy and simplicity of the approach.

Naïve Bayes Classifier:
In the field of Artificial Intelligence and Machine Learning, there are some classifiers which are from elementary "probabilistic classifiers" background. Naïve Bayes classifier is one of them and is based on the principles of Bayes' theorem with high independence Naïve assumptions among the features. It is one of the most elementary Bayesian models. Kernel density estimation can be coupled with them to achieve higher accuracy. Naive Bayes technique is one of the most elementary techniques to construct classifiers. The models that label problem instances with classes are called classifiers. The problem instances can be represented by a set of features where all the labels are selected from a finite set. The Naïve Bayes technique is not a standalone algorithm, but a set of algorithms which are based on a principle. It assumes that a specific feature's value is unconstrained of the values of any other features.

Naïve Bayes is one of the most useful techniques for filtering emails. Tackling spam email was started in the 90s and Naïve Bayes was hugely used in those methods.

This approach typically uses a bag-of-words feature to recognize spam-email, which is one of the most widely used approaches in text and document classification. This technique does the work by matching up the presence of features (typically words, sometimes sentences for greater accuracy), with non-spam and spam emails and then use Bayes' theorem to find the probability of that mail to be spam or non-spam email. In our proposed model to detect fake news, we have used an almost similar approach.

Experimental outcomes on datasets show that the method performs significantly well. An accuracy of 58% has been achieved for the Liar dataset which is least of all. The

Accuracy for the Fake News Corpus dataset is 81% and for Ban Fake News dataset it is 93%. So, Ban Fake News dataset outperforms the other two datasets.

Usually, fake news contains similar word sets that typically indicate the chances of news to be fake or non-fake. But it should be kept in mind that it's inappropriate to mention a report is fake since some word sets arise in it, however, these word sets hit the prospects of this certainty. The principal goal is to handle every single word of the news report separately. This section will introduce the approaches taken in previous attempts to the detection of the fake news. Firstly, we will introduce the papers that tried to solve this problem using the textual news content. The latter papers added information about the authors and spreaders into the account. The multi-modal approach included visual data as part of the predictors. Lastly, the computational fact checking takes acknowledge-based approach trying to use real-world knowledge. The survey gives a comprehending overview unifying all of the approaches lately used in this area.

2.1.1 Content-based fake news detection

Automatic Detection of Fake News [18] introduced two novel data sets.

The first data set being scraped genuine data from reliable news archives in 6
main domains and their fake equivalents were created by crowd sourcing
on AMT workers. The second one contained both real and fake news about celebrities gathered from the internet tabloids. The features in this study consisted of vector space representation of n grams in Term Frequency-Inverse Document Frequency (TF-IDF) metric and stylometric features in a form of punctuation from LIWC lexicon, psycholinguistics from LIWC lexicon, readability, and syntax. The detection was carried out by Support Vector Machine (SVM) classifier.

The results differed depending on the used dataset. For the primary Data set, the most effective performing classifier was supported stylometric features, Punctuation and Readability, while the other performed well with both, full LWIC lexicon and vector space representation features.

FND Net [9] uses Kaggle fake news data set on which it outperformed the implementations based on feature engineering and classical machine learning algorithms from the previous article. In this case, the model focuses solely on the vector space features where they used GloVe algorithm for embedding the words into the 100-dimensional vector used as an input for FND Net architecture. This deep learning architecture is based on adjusted Convolutional Neural Network (CNN) network in which they concatenate 3 concurrent convolutional layers and followed by dense layers. It outperformed both classical machine and deep learning using CNN and Long Short-Term Memory (LSTM) models. For the future, they suggest the use of multi-model word embedding’s.

Defending against Neural Fake News [20] proposes natural language generation model named Grover that can mimic the style of the real news. Similarly to the FND Net, it uses a neural network, however, with a different approach. It uses the Generative adversarial network that consists of two parts (players), generator and detector. The role of a generator is to create the best Fake News and the detector needs to detect it. When the generator improves, the detector improves as well. The neural Fake News created by architecture similar to Grove is difficult to detect by traditional methods, so good detectors will be crucial in the future. For the info, they generated the Real News dataset, an outsized corpus of reports articles from Common Crawl1.

Fake News Detection Using Deep Learning Techniques [21] compared Logistic Regression (LR), Naive Bayes (NB), SVM, Random Forest (RF) and Deep Neural Network (DNN) classifiers. They used usual text preprocessing from the natural language processing (NLP) domain (such as stemming, removal of stop words, etc.) to perform experiments on the LIAR dataset. As a result, they confirmed the findings of FNDNet that Deep Neural Networks outperform traditional machine learning methods.

Fake News Detection Using A Deep Neural Network [22] also compared various models using vector space representation as TF-IDF similar to the first article, but also Hasing Vectorizer. They implemented CNN, LSTM, NB, Decision tree, RF, and K-Nearest Neighbors (KNN).The performance of the algorithms decreased respectively to the mentioned order. The best performance was reached when combining CNN and LSTM confirming the findings of good performance of deep learning models. For the data, they used a combination of Kaggle datasets.

In the Text-mining-based Fake News Detection Using Ensemble Methods [13] paper they used vector space representation and stylometric Features similarly to Automatic Detection of Fake News using Ensemble methods. In this case, the stylometric features were divided into three separate feature subsets. The first one contained several unique words, complexity, Gunning-Fog index, character counts with and without whitespaces, Flesch-Kincaid readability score, etc. The second data set is presumed the lying detection dataset with specifications that may be separated into the subsequent categories: Quantity, Vocabulary, Grammar, Flesch-Kincaid score, Uncertainty, etc. The last subset consisted Of write print feature set for authorship attribution in short documents with categories: Character, Word, Syntactic, Structural, and Content. For vector space representation they used several options bag-of-words (BoW), TF-IDF, continuous bag-of-words (CBoW) that predicts word from the given context using neural network, skip-gram (SG) that also predicts next contextual word, both usingWord2Vec And Fast Text tools. Feature selection was used on both types of features. Stylometric features were selected by re-cursive elimination of weakest features. In word vector space features vocabulary was narrowed by lemmatization and stemming together with Chi-square tests for feature selection. As classifiers they used RF, NB (Gaussian and Multinomial), SVM, KNN, LR, bagging with general bagging classifier and extra trees classifier, and boosting with Ada Boost and Gradient boosting. The best overall accuracy was received by was by Gradient boosting using CBoWWord2Vec embedding's, outperforming all the non-ensemble machine learning algorithms. However, it is noteworthy that CBoW representation improved performance of non-ensemble algorithms. Using Rhetorical Structure Theory to Detect Fake Online Reviews [14] uses Rhetorical Structure Theory (RTS) mentioned in the journal above to detect fake data reviews. As a dataset, they used Deceptive Review (DeRev). They used groupings of certain RST features to form common macro-relations. By the analysis of the corpus, the fake reviews have more Elaboration, Joint and Background macro relations and the true reviews had more Evaluation, Contrast and Explanation, and only the true reviews contained Comparison macro-relations. This study shows that the authors paid to make fake reviews tend to deploy deceptive pragmatics seen in RTS. They violate the genre convention with tendencies to mention the title, author or content. Identification of corrupt news by the guidance of Sentiment Analysis [15] takes path of TF-IDF, sentiments and cosine similarity scores on NB, RF classifier trained on LIAR dataset. They claim that the use of sentimental score increases the accuracy of the model.

2.1.2 Authors and social media additional influence

Credibility-based Fake News Detection [23] contains the analysis of the influence of certain source and content measures on the credibility of the news. From the point of the source, the higher number of coauthors is a strong indicator of credibility, where anonymous information was mostly False. The graph of co-authorship shown that the authors of fake news are more likely to create articles together and vice versa. The same applies to the authors' affiliation the trust full organization and authors' history with the fake news. The content based features consisted of the aforementioned stylometric features consisting of sentiment analysis, domain expertise, argumentation, readability, character/word/ sentence Count, and presence of typos. After combining these two sources of credibility, they achieved a significant improvement in the accuracy of the final model by adding only three source-based features.

Neural User Response Generator: Fake News Detection with Collective User Intelligence [24] proposes Two-Level CNN with User Response Generator (TCNN-URG) is used to investigate the veracity of the news based on both content and the previous reactions of the users of the similar articles and generate their possible response to the new information. This method is useful for the early detection of fake news when the real user reactions are not available. They used Weibo dataset and used also their own Twitter dataset. User Response generator is based on Conditional Variation Auto encoder.
CSI: A Hybrid Deep Model for Fake News Detection [19] combines textual data with user response as the aforementioned article as well as the source information about the users that promoted the news. It consists of the three modules, the first one processes the text and Response using RNN network, the second analysis source information about the credibility source users and their groups and the last combines these approaches. Tested on Twitter and data sets.

It Proposes reinforcement learning for the next research. Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter [20] utilizes both vector space representation and linguistic measures as an input for a neural network that classifies the news whether it is a true News or one out of 4 levels of the fake news (satire, hoax, click bait or propaganda). The work done in [21] focuses on the credibility evaluation of the whole websites. It discusses the current state of the research in this field and recent drawbacks, e.g. cost of external APIs and shutdown of Google PageRank. Their research focuses on the web credibility model with the use of solely content-based features, disregarding user-based social features for the strong bias they introduce the final model confirmed by ANOVA test. The final model was evaluated on two data sets, Microsoft Dataset and Content Credibility Corpus, both consisting URLs and their credibility on the Linker 5-star Scale. The content based specification they utilized contained readability, PageRank information, General Inquirer (LIWC-like dictionary), Vader lexicon (sentiment), Lexical Categories (Lex Rank, LSA), Authority data (address, contact Email, etc.), social tags, whether the webpage is open-sourced, and their HTML2seq feature in a form of bag-of-tags (based on Bow). The Credibility prediction was done in two configurations classification and regression. As a result, they tested this model on the real world fact-checking problem where it proved to belittle web non-credible and support credible websites based on the supporting and opposing Claims.

2.1.3 Multi-modal approach

SAFE: Similarity-Aware Multi-Modal Fake News Detection [25] implements Multi-modal detection where it incorporates, both textual and visual content-based into the detection of the fake news. Even though this was previously used, their approach is novel in the way they process image data to text and take similarity between textual and visual data into consideration. As a baseline for their experiments they used LIWC mentioned above for the textual data, as well as VGG-19network for visual data, and at-RNN network for multi-modal data, all of which were outperformed. For future work, they propose to use additional information about the network and video.

2.1.4 Computational fact-checking

The Kauwa-Kaate Fake News Detection System: Demo [26] uses fact checking by querying against fact-checking websites. Supports querying text, images, and videos. It is not a fully automated system for Fake News detection but it supports fact-checkers. It uses information Retrieval approach when it queries the index of recently scraped articles from relevant sources. They are currently trying to establish and use a bias of the site on the certain domain using links and tweets to the site. They are querying images by using image matching library for signature strings. They handled screenshot by converting it to the text and querying it as a text. For the videos, they used signatures of the smaller scenes for matching.

2.1.5 Surveys

Fake News: A Survey of Research, Detection Methods, and Opportunities [26] is a survey that covers (i) methods to qualitatively and Quantitatively analyze, detect or intervene with fake news, (ii) four perspectives to study fake news based on knowledge, style, propagation, And credibility, (iii) news-related (e.g., headline, body text, creator, Publisher) and social-related information (e.g., comments, propagation Path and spreaders) used in fake news studies, (iv) techniques (e.g., feature-based or relational-based) used in fake news re-search, along with (v) review, classification, comparison and evaluation of current fake news and fake-news-related studies.

Automatic deception detection: Methods for finding fake news [31] provides a map to the current development to the veracity assessment methods, major classes and proposes hybrid system. Two major techniques are based on linguistic (stylometric and vector space representation) and network approach (defined in the meta-data/knowledge network). It proposes the use of data representation (Bow, frequencies, PoS, etc.), deep syntax (deeper syntactic structures – Probability Context-Free Grammars), a semantic analysis that extends n-gram plus syntactic approach with Profile compatibility, Rhetorical Structure and Discourse analysis (relations between linguistic elements). It also proposes the use of network approaches. These approaches complement well content-based linguistic approaches. They can be implemented in a form of extraction and examination of truthfulness based on factual knowledge about the real world. For example, querying DBpedia ontology and calculating the distance. The other possibility is to use Social Network Behavior to identify the credibility of the source.

2.2 Supervised learning

This thesis will focus on the use of supervised learning for prediction of the news credibility. The input of the supervised model consists of features (independent variable) describing each of the examples. The output of the supervised model is defined by the target (dependent) variable. Based on the data type of the target variable, we can divide supervised learning into classification (categories) and regression (Continuous variable). The training of the machine learning algorithm is a process, in which we present the machine learning algorithm with a training input-output pairs and the algorithm adjusts its parameters to most closely describe the training set by minimizing the output of the error/distance function. For evaluation of the algorithm, we use the testing set, which is a set previously unseen by the machine learning model. One of the major problems of supervised learning is over fitting that occurs when the model so fitted too close to the training data, as well as under fitting when the model is not fitted enough for the given task.

2.3 Performing query in encrypted data

AtutorialbyArasuetal.,2014 informs that querying encrypted data quite challenging, highlighting the obstacles in trusted hardware and performing encryption in the client- based encrypted. Li et al.,2014proposed a fast range query processing scheme for fake database that works against chosen keyword attack (INDCKA). The focused idea in this study is to maintain indexing elements in a full furnished binary tree called PB tree which satisfies structure in unique ability (i.e., two sets of data it ems have the same PB tree structure if and only if the two sets have the same number of data items) and node in distinguish ability (i.e., the values of PB tree nodes are completely random and have no statistical meaning). For efficient query processing, they used PB tree traversal width minimization and PB tree traversal depth minimization algorithm.

The worst case of complexity of the query checking algorithm using PB tree need O(R*log n) time, where n is the full number of data items and R is the set of data items in the query outputs.

Sahaetal.,2016 proposed an approach to secure sensitive fake data stored in the public cloud with the benefits of symmetric and asymmetric encryption keys using AES encryption algorithm. The authors proposed a framework for data-centric WSN application. Another objective of thesis to establish secure channel for data communication that could tackle attacks like MITM, DoS. Different segments of the news data are encrypted using different symmetric keys. These keys are distributed on demand basis to authorized users only. Before data transmission, data encrypted using AES symmetric key encryption algorithm. The scheme is made scalable by distributing static symmetric keys only to the legitimate users, with fine grained access control mechanism. Data integrity during transfer is achieved using SHA-1 hashing. Saha, Saha,2018proposed a methodology for protecting the privacy and Confidentiality one abusing sensitivity association, also providing efficient searching scheme using meta data based search. A secure and efficient data retrieval strategy was proposed by Saha, 2018.Here entire news relation fragmented into different segments according to sensitivity. The data is divided into clusters using correlation between news and authorized users to whom they are assigned. In the query evaluation phase, a co-relation metadata was used to validate news and the authorized-user combination. Fu et al.,2017 proposed a content-aware semantic search scheme for encrypted data. They used a graphical model, called conceptual graphs as a knowledge representation tool. Specifically they vectorized the plaintext to transform them into real valued vectors and performs query in the vector space. However the vectorizaion process scan lead to deterministic behavior, reducing the security standard. There lated works in the literature provides any directions for encrypted search But overlooks a unique challenge in the fake data domain. Often multiple users have to access the same attribute of a table, while having different read-write access control. Also different rows in the same table can have many such associations with different users. However, if the data is stored in encrypted form, the problem of sharing such encrypted data in the same table and performing query while maintaining data integrity is not well investigated. Typically, role based access management are used for ensuring only authenticated users get access other data Mitraetal. 2018, but such roles are defined in the database management system, not the application level, which is a requirement for sharing encrypted data.

2.4 Preserving data integrity from internal attacks

The thesis focuses on the block chain technology for preserving data integrity that recently has been popular as a verifiable storage for electronic health records. For example, Azaria et al.,2016; Kuo, Kim, and Ohno-Machado,2017; Liu,2016 presented various secure block chain storage for caption data. The first key benefit of block chain that It is a peer to-peer, decentralized data base management system. Therefore, block chain is suitable for applications where independently managed health care entities collaborate with one another. Secondly, DDBMSs support create, read, update, and delete functions, while block chain only supports create and read functions, i.e., it is very difficult to change the stored data. Thus, block chain is very much usable as an un changeable ledger to store critical information (e.g. Insurance claim records).Although block chain is based on distributed technology and thus do not suffer from single point of failure, it would be costly for DDBMS to achieve that high level of data Redundancy block chain does (i.e., each node has a whole copy of whole historical data records).Thus, block chain should be used only for storing data that are important and are modification sensitive.

In Muzammal, Qu, andNasrulin,2019,howablock chain can be used as are national data base was studied. Furthermore, block chain as a storage for data for auditing purposes was proposed in Azaria et al.,2016.However, block chain being a linear data structure, querying time in a block chain also increases linearly. This becomes a huge problem as the number of nodes in a block chain can be in thousands or millions in practical database for a large organization. However, fast querying in such large block chain that stores fake news information is not very well studied. The approaches most related to this work are done by Roehrs et al.,2019and Xu et al.,2017.Roehrs et al.,2019implemented a distributed architecture that distributes data servers and reintegrates when queried, which increases the average response time and availability. Their architecture is formed using a P2P network, where health records are organized into data blocks comprising a linked list and a distributed ledger of health data. Xu et al.,2017proposed a hierarchical data structure to enable efficient querying in a block chain that stores educational certificates.

The proposed approaches to minimize query time in block chain do not exploit the data characteristics present in a typical database. For example, in a information system, data records are often searched by time ranges to know about the history of the fake news. Also these types of information are actually the prime candidates to be altered by malicious users. In the proposed approach, a hybrid security model is used where only hash values of such modification-sensitive data are store Dina block chain to reduce overhead and the actual data is stored in encrypted form in a Cassandra database. Also, the block chain storage uses a hierarchical data structure using the timestamp information present in the fake data that supports a much faster query.

Chapter 3

Background

3.1 Security in cloud database

The cloud computing model is transfers computing infrastructure and data to third- party service providers that manage the hardware and software resources which enables on-demand, anytime-anywhere access and cost reductions. Many organizations have started shifted electronic information to the cloud storage. Not only it simplifies the exchange records between the other participating organizations, but also makes the cloud a record storage center which is permanent and remotely accessible by many users. The fake data stored in cloud makes the treatment better by retrieving data history from the database before going and get to know about the health issues of the fake data. However, maintaining the confidentiality of the data stored in cloud is a major issue.

The data stored in such a cloud data base can be stole

Nina number of ways:

1. Data stolen while transferring using network
2. Direct abuse by untrusted cloud service provider
3. Third party attack in the network or on the cloud service provider
4. Legal issues: Law in the nation of the cloud service provider may enforce it to reveal all the data.

One way to implement security of the data in a cloud database is to encrypt the data before sending to the cloud this involves sharing a key with the cloud provider. In this method all the encrypted data sent by client are first decrypted with the key and decrypted data is stored in the DBMS. Now the client application can send an encrypted query which is decrypted and run on the server DBMS. To ensure confidentiality the result of the query is encrypted by the security module in the server. The client also having the key can decrypt this encrypted result and view the data. An over view of the scheme with a trusted cloud provider is shown inFig.3.1. Al though this simple solution works well, it can be only implemented with trusted and secure cloud provider. Also there is still a risk of leakage of sensitive data because

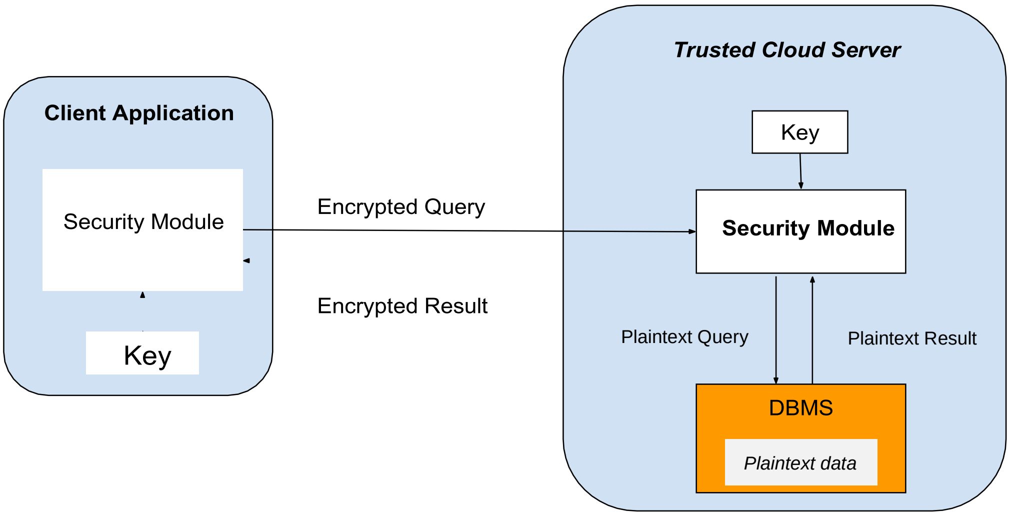

FIGURE 3.1: Data transfer in trusted cloud provider.

Securing a server in the cloud is not easy and there is a chance of third party attack. Also legal issues can cause loss of confidentiality.

In an alternative approach called the *data-at-rest encryption*, no key is shared with the server and only encrypted data is stored in the cloud DBMS, shown in figure3.2. Client application sends encrypted data to the cloud, and the cloud provider directly stores the encrypted data in the DBMS. In this case, the cloud provider can never see the actual data. So any attack on the cloud server, or an untrusted cloud provider only gets the encrypted data which is of no use without the decryption key. Also this data is safe from legal issues.

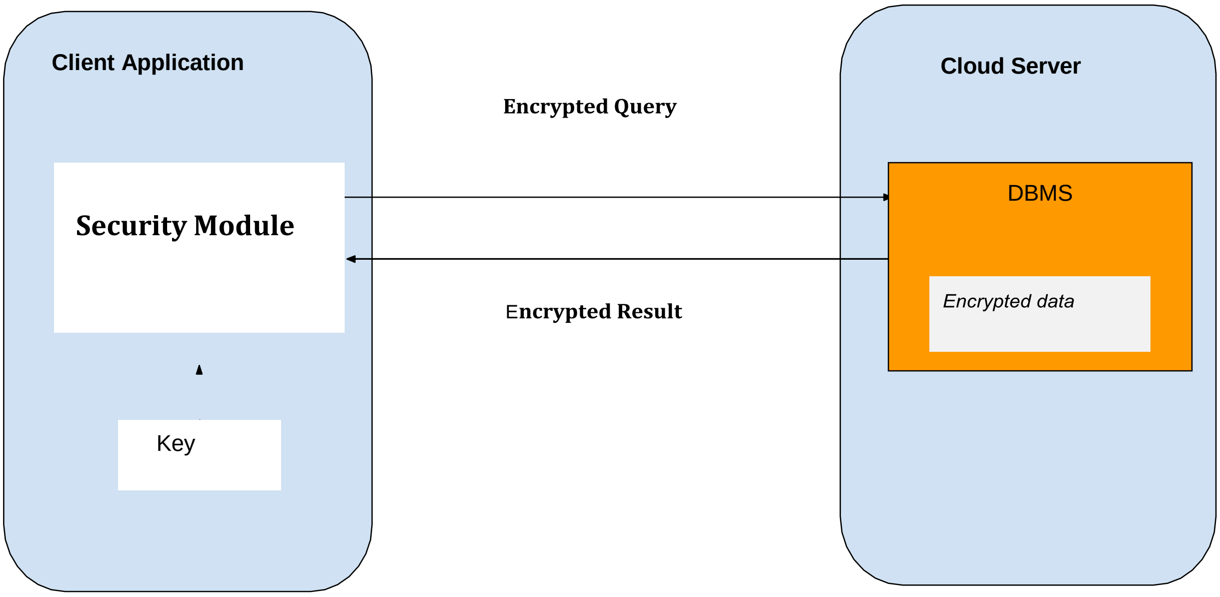

FIGURE 3.2: Data-at-rest encryption in DBMS.

Even though this mechanism conforms the required security standards, there are several issues with storing encrypted data in the database management system, as re- ported by Arasu et al.,2014:

1. Performing search query is difficult on encrypted data. This is because the user is expected to provide a query containing attributes in plaintext, which are to be matched with encrypted data.
2. Some queries require data to be ordered by some column value. Since a primary requirement of any encryption algorithm that there should be no relation between the character soft he plain text and the cipher text, i.e. without the key the plain text cannot be predicted from the cipher text. So obviously, the order of some data records (plaintext) determined alphabetically or numerically will not be the same

if the records are encrypted. So performing such queries where the actual ordering of the data is required is challenging in a encrypted data base.

3. Another set of queries like sum, average etc. require mathematical operation to be perform on the data records. The encryption scheme that allows performing a mathematical operation on encrypted data and getting the same result as if perform on actual data is called Homomorphism encryption. This is challenging and there are only a few Homomorphism encryption schemes available, that to allows only a restricted set of operations.

3.2 Cassandra

In this section the motivation for using Cassandra as the database is provided, along with a brief description of its security aspects.

3.2.1 Introduction to Cassandra

Cass and raise a highly scalable, distributed, structured key value store No SQL data base that is chosen to demonstrate the security model presented in the thesis. It is a mixture of the distributed system technology like the Dynamo database and the data model is column family based like Google's Big Table. Cassandra is eventually consistent; i.e.it may not be consistent all the time but will be consistent eventually after a finite time. Cassandra supports four types of storage model: Wide Column Store/Column Families, Document store, Key Value/Tuple Store,

Eventually Consistence Key Value store, Graph Data bases. The design goal of Cassandra is to handle large amount of work load across multiple nodes without single point of failure by replicating data. All the nodes in a cluster act the same role. Each independent node interconnected to other nodes at the sometime.

Cassandra deals with unstructured data and flexible schema. Internal data model of Cassandra and the equivalent relational data base terminologies are shown in Table 3.1.Attributes are only blob-values in the internal model and the attributes of one row are always stored sorted by the name of the attribute, when written to the internal storage called the *SS Table*, shown in Figure 3.3).

TABLE 3.1: Terminologies of Relational DBMS and Cassandra.

Relational Model	**Cassandra Model**
Database	Key space
Table	Column Family(CF)
Primary Key	Row Key
Column Name	Column Name/Key
Column Value	Column Value

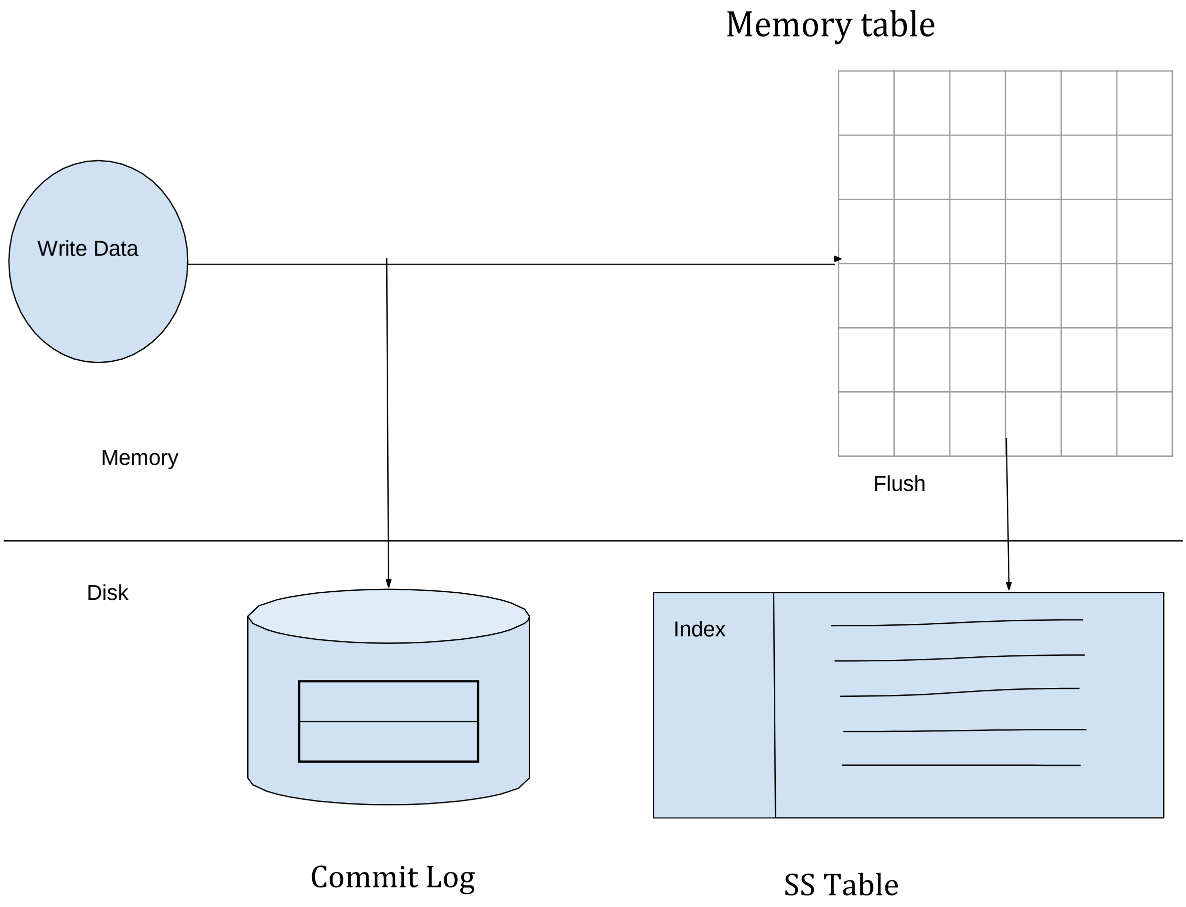

FIGURE 3.3: Data storage model of Cassandra.

Every tuple also contains one special attribute: the primary key, which is used for distributing and replicating the tuples across the nodes of the database cluster, in addition to its traditional use for indexing. Cassandra does not repeat the names of the columns in memory or in the SS Table. In the SS Table on disk, Cassandra stores data after flushing the *mutable*. Any data written to Cassandra is first written to a commit log before being written to a mutable. This creates durability in the matter of unreal shut down. On starting Cassandra any mutation sin the commit log is be applied to mutable.

3.2.2 Querying in Cassandra

Cassandra provides its own query language called Cassandra Query Language (CQL), available in the online documentation *Cassandra Query Language*. CQL is somewhat similar to Structured Query Language (SQL) that offers an easy interface to interact with the database. This is because the Thrift API, that was being used prior to CQL, was found difficult to comprehend. CQL supports various set of data type i.e native types, collection types, user-defined types, tuple types and custom types. CQL uses database roles to represent users or group of users.

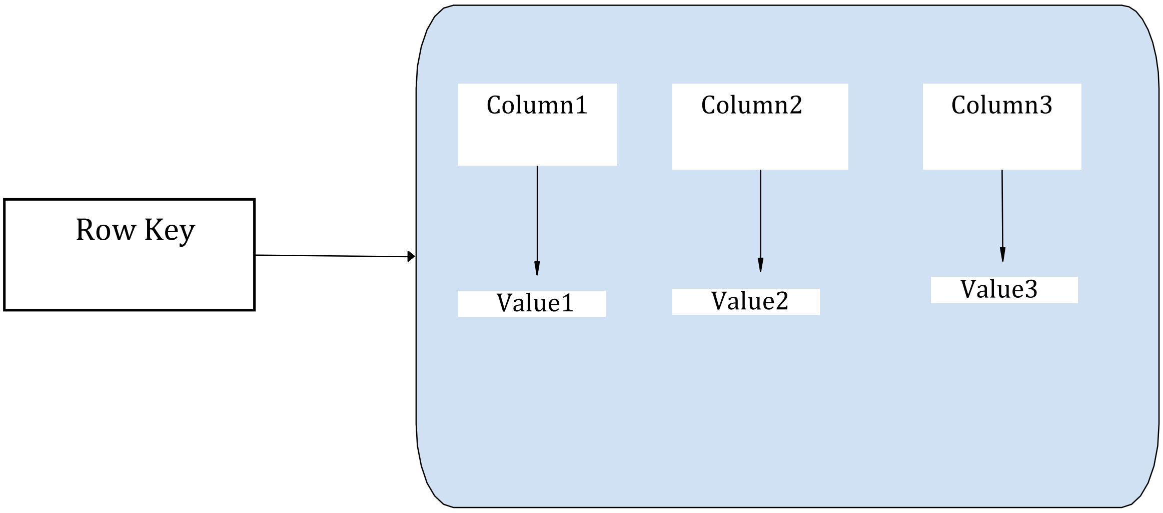

FIGURE 3.4: Cassandra look-up map using row and column keys.

CQL uses a map for efficient key look-up, and the its sorted nature gives efficient scans. In Cassandra, both the row keys and column keys to do efficient look-ups and range scans, as shown in Figure 3.4. The number of column keys is unbounded, In other words, i.e very wide rows are supported.

The difference between the internal data model used by Cassandra and the CQL data model is that the columns have fixed type sin the CQL data model, i.e. they are no longer just blob values, and the primary key can be a compound key, i.e. consisting of various columns. Additionally, the primary key is further divided into two different parts: the partitioning primary key and the clustering primary key, both of which can be made of an arbitrary amount of columns. The partitioning primary key is the one that is used as the primary key in the internal data model, and thus it is used to share and replicate the tuples. The clustering primary key is used when transporting CQL tuples into internal tuples to define their ordering, i.e. multiple CQL tuples with common partitioning primary key will be sorted in one internal tuple based on their clustering primary key, when written to the internal storage.

3.2.3 Security aspects of Cassandra

There are three major components to the security features provided in Cass and ran by default: Internal authentication, internal authorization and SSL/TLS encryption for client and internode communication. A detailed documentation on the security provided by Cassandra is available online *Cassandra Security Documentation*.

Internal Authentication

Internal corroboration is based on Cassandra-controlled roles and passwords. Role- based authentication encompasses both users and roles, i.e. different users can have different roles in the database and the specific roles are granted using password based authentication. The roles can represent either actual individual users or roles that those users have in administering and accessing the Cass and ran cluster.

Internal Authorization

Cassandra supports object permissions that are assigned using Cassandra's internal authorization mechanism for the following objects: key space, table, function, aggregation, role sand MBeans (available only in Cassandra 3.6 and later).Authenticated roles with passwords stored in Cassandra are authorized for selective access. The permissions are stored in Cass a datable.

Permission is configurable for CQL commands that are used to create or manipulate data. The allowable commands are: *CREATE, ALTER, DROP, SELECT, MODIFY, and DESCRIBE*, which are used for the generic interaction with the data base. The *EXE- CUTE* command may be used to grant permission to a role for the *SELECT, INSERT, and UPDATE* commands. Additionally, the *AUTHORIZE* command can be used to grant permission for a role to *GRANT, REVOKE or AUTHORIZE* o the role's permissions.

Nodes. However, if Cassandra itself runs in a cloud computing environment, then the stored data can be compromised. To store data in an untrusted cloud provider, the thesis presents a secure application layer on top of Cassandra.

Chapter 4

Proposed Method

4.1 Overview

In this section, an efficient and secure system model is introduced for managing data stored in a public cloud environment. In short, this model provides a secure interface to the data that is stored in a distributed database, by tackling threats of both internal and external attackers. This model uses encryption to store data securely and thereby protecting its confidentiality. A secure interface is provided to news for accessing the required data on-demand for both read and write requests. Specifically, the model allows encrypted data to be shared by several concerned users in a secure manner; i.e. it provides seamless access to the encrypted data for permitted users, while limiting the access for other users in the system along with the internal or external attackers.

Another part of this work to prevent unauthorized modification of data by internal attackers block chain has been used. Internal attackers may be able to modify the data in the database. To prevent this, hash values of each data record are stored in a block chain. These hash values of the data cannot be modified due to the immutable property of block chain. This means, once hash values of data records has been writ- ten to a block chain, not even an authorized user can change it. However, this can be only applied to be data that cannot be modified. In the data context, there are much such information that need not be changed after inserting in the database.

4.2 Approach

The proposed approach of the secure system model consists of different modules as described below. A details of the system architecture is may be viewed in Figure 4.1. In the following sub-sections different modules of the system are described.

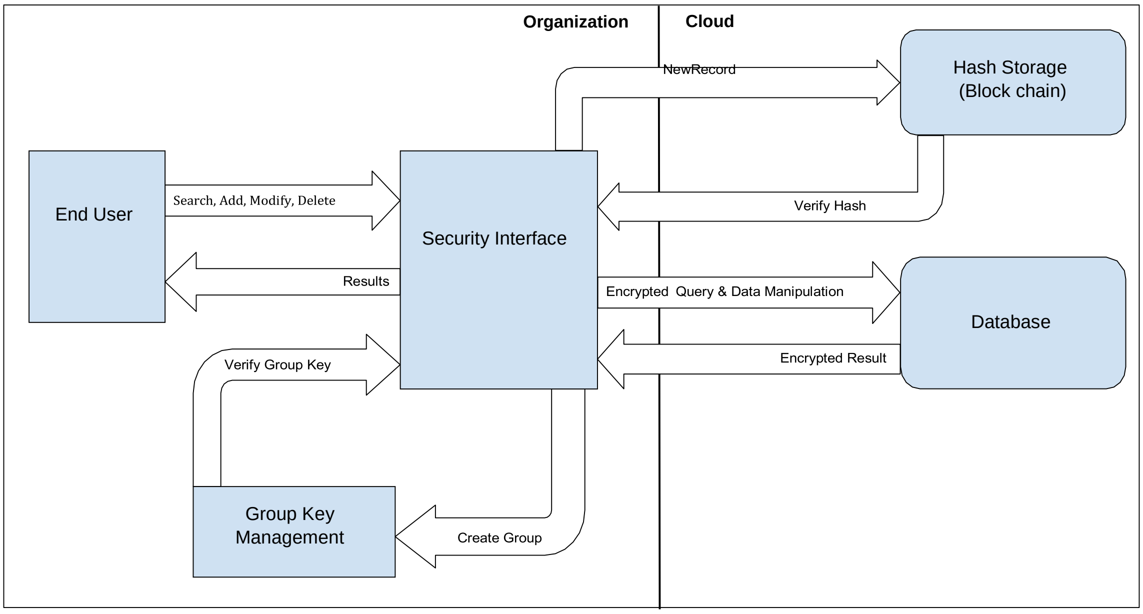

FIGURE 4.1: An overview of the system model.

4.2.1 Block chain storage

Although storing data in encrypted form prevents attacks from external threats, internal attackers who are maliciously logged in to the system as authorized users, can still modify sensitive data. An approach to maintain data integrity is proposed in this section that uses a *block chain* to check and prevent unauthorized data modification.

A block chain is a list of records, called *blocks*; such that each block contains a cryptographic hash of the previous block, a timestamp and some arbitrary data. We also propose an efficient strategy to search in the block chain that is applicable to data base systems. Specifically, has h values of each sensitive data

record that are in the encrypted form are stored in a block chain. These hash values of the encrypted data cannot be modified due to the immutable property of a block chain. This means, once has h values of encrypted data records has been written to a block chain, note venin authorized user can change it. Also, because of the immutable property of a block chain, only the parts of the data that's hold not be modified are stored in the block chain.

In the data context, there is much such information that needs not be changed after inserting in the database, news but some of the information needs to append according to news going on such as

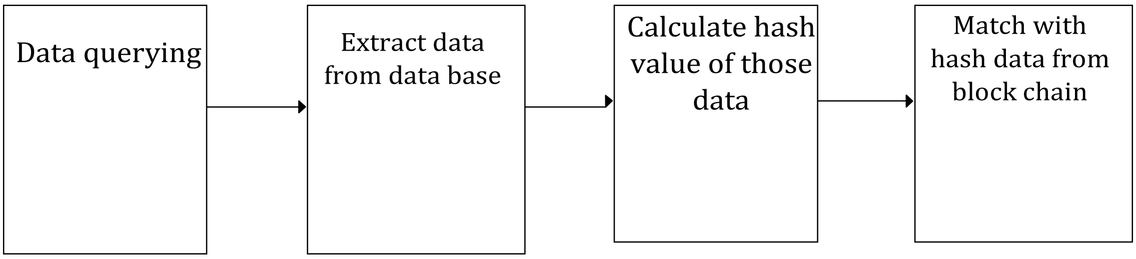
Data querying
Extract data from data base
Calculate hash value of those data
Match with hash data from block chain

FIGURE 4.4: Workflow to detect unauthorized data modification.

Data in a block can't be modified even by authorized user, because this wills invalidate the internal hashes of the block chain that are dependent on previous blocks, i.e. the hashes are calculated in cascading manner. So, even if an authorized user changes some parts of the data in the encrypted database, the authorized user can't store the changes in block chain. Data is always inserted in the last blockier.

The block chain only supports insert and select queries, no update queries are allowed. The data integrity check is performed during a select query. So if in the meantime, if a malicious user modifies a data record in the encrypted data base with a valid Key, still the data alteration can be check reducing the block chain. The main motivation for using the block chain to store the hash values is that, it is very difficult to change the data stored in the block chain, so the has h values are protected from the internal attackers. Also, no data base user has access to the block chain. The data in the block chain is inserted and queried internally by the application. This prevents unauthorized users from forcefully tampering the block chain itself.

4.2.2 Efficient Query in block chain

The block chain storage is used for storing hash values of the modification-sensitive parts of the news data. However, as block chain is essentially a linked list of blocks, searching in the block chain for has h validation must be performed sequentially, i.e. random access to a block in the block chain is not permitted. In practice, an organization can have hundreds of news getting admitted per day. This can result asymptotically, the time complexity is O (n) in the worst case, where n is the number in sequential search in a block chain with millions of blocks, which is very inefficient. Of blocks in the block chain. The data in the block is always time stamped. We exploit this property to implement the block chain in a hierarchical manner to support efficient the topmost level of this hierarchical data structure consists of a list ordered by year, which is extracted from the timestamp. The middle level is a list of lists, where each list consists of months that belong to the year, determined by the position in the outer list. Similarly, the lowest level consists of the actual block chain, stored as a nested list, where individual lists contain the blocks for a month. The procedure for validating the data integrity of news is described below:

1. First, the encrypted data record is fetched from the database, which includes the same timestamp stored in the block chain.
2. Hash value of the encrypted data record is calculated using SHA-256, that yields H_1
3. The timestamp is partitioned into year and months.
4. Then the top level list is searched using the year as search key.
5. Using the returned position and the month, the mid-level nested list is searched and soon.
6. At the lowest level, data is searched in a linear fashion, and the corresponding data block is extracted.
7. From the extracted data block, the value of the encrypted data record is calculated using SHA-256, that yieldsH_2.
8. If H1= H2 the data is returned, else an error is raised to show that data integrity is lost.

In the worst case, total number of comparisons is k+l+m, where k is the length of the top

Level list, l is for the mid-level and m is for the bottom level list. As the number of months in a year is constant number, the total asymptotic complexity is given by O (k + m), where k + m is much lesser

Chapter 5

Implementation and Evaluation

In this chapter the details of the implementation of a system that follows the security model proposed in Chapter 4 are described.

5.1 Implemented System

The proposed work is implemented using the python programming language, using Cassandra as the database. The CQL commands for using the Cassandra database are presented below:

- **Creating the key space:**
  CREATE KEYSPACE *user data base* WITH
  REPLICATION= *{'class':' Simple Strategy', 'replication_factor':'3'}*
  AND DURABLE_WRITE= *true*;

The software components of the system and their versions are shown in Table5.1.

TABLE 5.1: Versions of the software components of the implemented system.

Component type	Name	Version
Programming environment	Python	3.6.5
Database	Cassandra	3.11.3
Query engine	Cqlsh	5.0.1
Operating system	Ubuntu	14.04

For performing encryption, hashing and for connecting to the

Cassandra database several python libraries were used. Table 5.2 shows the details of the libraries.

TABLE 5.2: Details of the python libraries used.

Library	**Purpose**	Version
Cassandra-driver	The driver connects to the Cassandra Cluster from python. The driver also supports performing CQL queries.	3.13
Pycryptodome	Performs encryption and decryption Using AES algorithm.	3.5.1
Bcrypt	Performs hashing using SHA-256 algorithm.	3.1.4
Pillow	Performs image manipulation	5.0.0
Pickle	Performs python object serialization, Used to store block chain objects as files.	11.0.1

5.2 Results

The system was experimented with using two different machines, both running Ubuntu14.04 operating system. Table 5.3shows the hardware specification for these two machines. One of the machines was used as a cloud server and another was used to run the client application. The two machines were connected by a personal Wi Fi network with a link speed of 150 megabytes per second.

TABLE 5.3: Hardware specifications of experimented machines.

Machine	**CPU speed**	**Memory (Speed)**	**Disk (Speed)**
Server machine	2.7 GHz, 4cores	4 GB (2400 MHz)	1 TB (5400 rpm)
Client machine	2.2 GHz, 2cores	4 GB (1600 MHz)	512 GB (5400 rpm)

In the following sections the experiments with the proposed system, performed in these machines are described and the results are reported.

5.2.1 Evaluation of encrypted search

1. The cipher text of the search term is stored in a metadata table, along with the metadata generated for the user. When the data is to be queried, the metadata is used as a secondary index, to first find the cipher text of the search term. Then the cipher text of the search term is used to find the actual data row.

Table 5.4shows the comparison of the above mentioned system variants. The results were calculated by averaging the measured time of 100 queries in a database with 5000 inserted rows.

TABLE 5.4: Average query times for searching in a non-deterministically encrypted database.

Query type	Metadata used	Search term	Time
Select	Yes	Encrypted	270.1 ms
Select	No	Plaintext	13.6 ms

Note that the Meta data-less method performs query faster because it does very minimal processing. But the security of this method can be compromised as the search keys cannot be stored in encrypted form. Whereas, the query using metadata takes longer, as it has to perform one extra query to get the metadata-key pair. But, as all the information in the actual table is encrypted, this method is much more secure. Even though the time taken for the metadata based search takes longer time, it does not affect the experience, as the time is within seconds, i.e. it performs the query in real time.

5.2.2 Evaluation of block chain search

Detailed experiments were performed to compare the naive linear data structure and the proposed hierarchical data structure approach in a block chain, by vary i g the number of blocks and by measuring both the insertion and the query times. Table 5.5 shows the results of the comparison.

TABLE 5.5: Query time comparison between linear and the proposed hierarchical data structures.

Types	No. of blocks	Insertion time(ms)	Query time (ms)
Linear	2052	0.001515	0.202277
Hierarchical	2052	0.010424	0.005797
Linear	11172	0.005885	5.996496
Hierarchical	11172	0.046613	0.027285
Linear	45372	0.018346	127.751428
Hierarchical	45372	0.114271	0.091386
Linear	113772	0.044489	707.559454
Hierarchical	113772	0.294859	0.198096

From analysis of the results, it can be seen that insertion time for the linear data structure is always less than the proposed approach, but the difference between the measured times of the two approaches is not large. This is because in the proposed approach many calculation sand insertions are performed for the different levels of the data structure. Also, it is import ant to consider that the number of insertion operations in a health-care organization is much less than the search operations. So, the small increase in the insertion time in our approach can be traded off for much faster searching performance.

Whereas, the proposed approach always stakes much lesser time that the line alone. Also, as the size of the block chain increases, the proposed approach becomes more and faster than the linear approach

Chapter 6

Concluding Remarks and Future Directions

6.1 Conclusion

Fake News is becoming a threat to the community and affecting the lives of common people. There are many good approaches in fake news detection in English language but the researches in Bengali language and most other regional languages is in a very nascent stage. In this study, we are using the Naïve Bayes to be the classifier which can distinguish between Fake and Non-Fake News. This elementary AI algorithm has shown this much great result on such a significant problem as the detection of fake news. The outcomes of this experiment suggest that AI techniques and algorithms may take a crucial role in tackling this kind of problems which are driving the world insane.

Deep Learning using Neural Networks are proven to increase accuracy and efficiency in such kind of experiments. In this work, machine learning techniques are used. An accuracy of 82% in English language and 93% in Bengali language are achieved. We wish to apply deep learning method in future.

The experiment has been done on two languages which are English and Bengali and has achieved a significant result. The same work can be done in other languages especially Indian languages like Hindi, Tamil and Marathi etc. As this experiment is independent of languages and independent of grammatical syntaxes it should work in other languages. So, the experiment of applying this algorithm in fake news detection in other languages is our next target in future.

Detecting fake image, video and stopping them before spreading over internet through social media, is another possible enhancement of this work.

No SQ Land distributed data base such as Cassandra is a good choice for s to ring data of large organizations. Cassandra clearly have advantages like unstructured data model and high scalability. However, Cassandra only provides security for client to data base node or inters node communication. But there are several security vulnerabilities that can lead to neither stole nor tampered data. This thesis presented an approach to protect the confidentiality of sensitive data of a news store Dina cloud data base. The approach provides the directions to store data in a non-deterministically encrypted form and to search in such a encrypted database. The proposed approach tackles the problem of sharing encrypted data by different groups of user having different roles and permissions. Also, an approach to provide data integrity in secure block chain storage, and a method to efficiently search in such a block chain, exploiting the properties of data is presented. The proposed method uses a hierarchical data structure that provides an asymptotic lower bound than a linear search. The proposed system provides real time query performance, while conforming to a high security standard.

6.2 Future Work

There are many scopes of future research from the work presented in this thesis. In its current form, the proposed system is limited to simple select queries only. How to extend the proposed approach to support more complex queries can be investigated in future. Also, if a health-care organization has to continuously monitor news and securely store such real times teaming data, sharing of such encrypted, real-time data by different user groups can be of interest. The proposed system provides same thought validate data integrity in a secure manner, but in case of un-authorized modification, restoring the original data can also be a challenge.

Bibliography

4. References

1. Tavernisen, S.: As fake news spreads lies, more readers shrug at the truth. New York Times, 6 December 2016. http://nyti.ms/2lw56HN

2. Vlachos, A., Riedel, S.: Identification and verification of simple claims about statistical properties. In: 20th Conference on Empirical Methods in Natural Language Processing (EMNLP 2015), pp. 2596–2601, September 2015. doi:10.18653/v1/d151312

3. Acemoglu, D., Ozdaglar, A., Parandeh Gheibi, A.: Spread of (mis) information in social networks. Games Econ. Behav. **70**(2), 194–227 (2010)

   CrossRef MATH MathSciNet Google Scholar

4. Afroz, S., Brennan, M., Greenstadt, R.: Detecting hoaxes, frauds, and deception in writing style online. In: 33rd IEEE Symposium on Security and Privacy (SP 2012), pp. 461–475. IEEE, May 2012

   Google Scholar

5. Joachims, T.: Text categorization with Support Vector Machines: Learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998). doi:10.1007/BFb0026683

   CrossRef Google Scholar

6. Wang, S., Manning, C.D.: Baselines and bigrams: Simple, good sentiment and topic classification. In: 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012), pp. 90–94, July 2012

   Google Scholar

7. Tang, D., Qin, B., Liu, T.: Document modeling with gated recurrent neural network for sentiment classification. In: 20th Conference on Empirical Methods in Natural Language Processing (EMNLP 2015), pp. 1422–1432, September 2015

Google Scholar

8. Frege, G.: Sense and reference. Philos. Rev. **57**(3), 209–230 (1948)

CrossRef Google Scholar

9. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations (ICLR 2015), May 2015

Google Scholar

10. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2016), pp. 1480–1489, June 2016

Google Scholar

11. Horne, B., Adali, S.: This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In: Workshop of the 11th International AAAI Conference on Web and Social Media (ICWSM 2017), May 2017

Google Scholar

12. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. In: 19th Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), pp. 1724–1734, October 2014

Google Scholar

13. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: 19th Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), pp. 1532–1543, October 2014

Google Scholar

14. Gillin, J.: Politifact's guide to fake news websites and what they peddle. PunditFact, 20 April 2017. http://bit.ly/2pHYKDV

15. Glader, P.: 10 journalism brands where you find real facts rather than alternative facts. Forbes, 1 February 2017. http://bit.ly/2sXPpvf

16. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), pp. 55–60, June 2014

Google Scholar

17. Kim, Y.: Convolutional neural networks for sentence classification. In: 19th Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), pp. 1746–1751, October 2014

Google Scholar

18. PÉREZ-ROSAS, Verónica; KLEINBERG, Bennett; LEFEVRE, Alexandra; MIHALCEA, Rada. Automatic Detection of Fake News. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association

19. KALIYAR, Rohit; GOSWAMI, Anurag; NARANG, Pratik; SINHA, Soumendu. FNDNet- A Deep Convolutional Neural Network for

Fake News Detection. Cognitive Systems Research.2020, vol. 61. Available from DOI: 10.1016/j.cogsys.2019.12.005.

20. ZELLERS, Rowan; HOLTZMAN, Ari; RASHKIN, Hannah; BISK, Yonatan; FARHADI, Ali; ROESNER, Franziska; CHOI, Yejin. Defending Against Neural Fake News. CoRR. 2019,vol. abs/1905.12616. Available from arXiv: 1905.12616.

21. HIRAMATH, C. K.; DESHPANDE, G. C. Fake News Detection Using Deep Learning Techniques. In: 2019 1st International Conference on Advances in Information Technology (ICAIT). 2019, pp. 411–415.

22. KALIYAR, R. K. Fake News Detection Using A Deep Neural Network. In: 2018 4th International Conference on Computing Communication and Automation (ICCCA). 2018, pp. 1–7.

23. SITAULA, Niraj; MOHAN, Chilukuri K.; GRYGIEL, Jennifer;ZHOU, Xinyi; ZAFARANI, Reza. Credibility-based Fake News Detection.2019. Available from arXiv: 1911.00643 [cs.CL].

24.QIAN, Feng; GONG, Chengyue; SHARMA, Karishma; LIU, Yan.Neural User Response Generator: Fake News Detection withCollective User Intelligence. In: Proceedings of the Twenty-SeventhInternational Joint Conference on Artificial Intelligence, IJCAI-18. InternationalJoint Conferences on Artificial Intelligence Organization,2018, pp. 3834–3840. Available from DOI: 10.24963/ijcai.2018/533.

25. ZHOU, Xinyi; WU, Jindi; ZAFARANI, Reza. SAFE: Similarity-Aware Multi-Modal Fake News Detection. 2020. Available from arXiv: 2003.04981 [cs.CL].

26. BAGADE, Abhishek; PALE, Ashwini; SHETH, Shreyans; AGARWAL, Megha; CHAKRABARTI, Soumen; CHEBROLU, Kameswari; SUDARSHAN, S. The Kauwa-Kaate Fake News Detection System: Demo. In: 2020, pp. 302–306. Available from DOI: 10.1145/3371158.3371402.