Sentiment Analysis on Hotel Reviews using Machine Learning

Skill Based Lab-I mini-project report submitted in

partial fulfilment of the requirements of the degree

of

Computer Engineering

by

Swijel Peter Dmello   PID No. (EP1241011)

Under the guidance of

Ms. Shraddha More
Assistant Professor

Department of Computer Engineering

St. John College of Engineering and Management, Palghar

University of Mumbai

2024–2025

CERTIFICATE

This is to certify that the Skill based Lab-I entitled **“Sentiment Analysis on Hotel Reviews using Machine Learning”** is a bonafide work of **“Swijel Peter Dmello” (EP1241011)** submitted to University of Mumbai in partial fulfilment of the requirement for the award of the degree of **“Master of Engineering”** in **“Computer Engineering”** during the academic year 2024-2025.

Ms. Shraddha More

Project Guide

Dr. Nilesh T. Deotale	**Dr. Kamal Shah**
Head of Department	Principal

Skill Based Lab-I Report Approval

This project report entitled ***Sentiment Analysis on Hotel Reviews using Machine Learning*** by ***Swijel Peter Dmello*** is approved for the degree of ***Master of Engineering*** in ***Computer Engineering*** from ***University of Mumbai***.

Examiners

1.---

2.---

Date:

Place:

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Signature
Swijel Peter Dmello (EP1241011)

Date:

Abstract

As the trend to shop online is increasing day by day and more people are interested in buying the products of their need from the online stores. This type of shopping does not take a lot of time for a customer. Customer goes to an online store, searches for the item of his/her need and places the order. But, the thing by which people face difficulty in buying the products from online stores is the bad quality of the product. Customers place the order only by looking at the rating and by reading the reviews related to the particular product. Such comments of other people are the source of satisfaction for the new product buyer. Here, it may be possible that the single negative review changes the angle of the customer not to buy that product. In this situation, it might be possible that this one review is fake. So, in order to remove this type of fake reviews and provide the users with the original reviews and rating related to the products, we proposed a Fake Product Review Monitoring System. The fake reviews dataset has been used to detect false positive and false negative reviews. We propose an approach to detecting fake reviews through various advanced machine learning techniques like Support Vector Machines, Decision Trees, etc. The system put forward is a web-based solution that provides an accurate result whether the given review is valid or not.

Keywords—*Sentiment Analysis, SVM, KNN, Decision Tree, Logistic Regression.*

Table of Contents

List of Figures

List of Tables

List of Abbreviations

SJCEM	St. John College of Engineering and Management
SVM	Support Vector Machine
SVC	Support Vector Classifier
KNN	K-Nearest Neighbors
DT	Decision Tree
LR	LogisticRegression
ML	Machine Learning
CSV	Comma Separated Value
DFD	Data Flow Diagram

Chapter 1

Introduction

The Social Web has made an intense change in the existence of everybody these days through communicating their perspectives on the web. The size of the client-created content is developing quickly. E-Commerce gives a polished experience to web-based clients. The sites permit their clients to give their criticism as surveys on their destinations. Over 90% of shoppers before buying any item or utilizing any assistance, it has turned into a propensity for perusing web surveys for a dynamic reason. Around 40% to 70% of audits given in Online locales are found as false surveys. The new clients as well as the current clients think about the surveys as their significant wellspring of data. Every one of the reviews composed is false. A portion of the audits are spam as they are composed for certain advantages like advertisement for their item, promotion of their item or administration, essentially spreading information or now and then out of these phony surveys might even get Financial Profits. Hardly any business ventures utilize an individual of letters to draw up manufacturing positive surveys on their items and corrupt negative audits on their adversary's items. Thus, simply accepting these internet-based surveys and settling on choices might turn out badly because not all reviews are authentic. It deludes clients and the absence of control in spam data spreading. Dimensionality decrease should be done to build the outcome. The Review Dataset is profoundly unique and veracity in nature as it is brimming with text information.

1.1 Motivation

Nowadays, the World Wide Web has drastically changed the way of sharing opinions. Online reviews are comments, tweets, posts, opinions on different online platforms like review sites, news sites, e-commerce sites or any other social networking sites. Sharing reviews is one of the ways to write a review about services or products. Reviews are considered as an individual’s personal thought or experience about products or services. Customer analyzes available reviews and makes a decision whether to purchase the product or not. Therefore online reviews are a valuable source of information about customer opinions [1]. Fake or spam review refers to any unsolicited and irrelevant information about the product or service. Spammer writes fake reviews about the competitors’ products and promotes their own products. The reviews written by spammers are known as fake reviews or spam reviews. Thus fake reviews detection has become a critical issue for customers to make better decisions on products trustworthy as well as the vendors to make their purchase.

The rising dependence on web-based surveys, especially in the friendliness business, has prompted a flood in the quantity of phony or deceiving surveys that can essentially influence a lodging's standing and a client's dynamic cycle. Counterfeit audits, whether positive or negative, are frequently made by contenders, promoting firms, or other vindictive entertainers to delude possible clients. This represents a basic test for both lodging the executives and purchasers who depend on these surveys to settle on informed choices.

1.2 Problem Statement

The target of this task is to foster a mechanized framework fit for recognizing counterfeit surveys from a lodging surveys dataset. The framework will use a few AI (ML) calculations, including Choice Tree, K-Closest Neighbors (KNN), Strategic Relapse, and Backing Vector Machines (SVM), to characterize surveys as either certifiable or counterfeit. The essential objective is to preprocess the dataset, extricate applicable highlights, and train the calculations to distinguish designs characteristic of phony audits. By assessing the presentation of these calculations, the venture plans to figure out which model yields the best precision in identifying counterfeit audits.

The project aims to preprocess the hotel review dataset by addressing missing values, text cleaning and tokenization. Key features such as sentiment, review length, and word usage will be extracted for training purposes. Machine learning models will be assessed based on accuracy, precision, recall, and F1-score to determine the most effective model for detecting fake reviews.

1.3 Objectives

The objectives are as follows:

- **To get** a hotel reviews dataset that contains labeled reviews (genuine or fake) for model training and testing.
- **To perform** data preprocessing using text processing techniques like tokenization, cleaning, and transforming raw review data into a usable format.
- **To extract** relevant features from the dataset, such as sentiment, review length, and keyword patterns, to train machine learning models.
- **To perform** feature selection and engineering to enhance the model's ability to differentiate between genuine and fake reviews.
- **To implement** machine learning algorithms, including Decision Tree, KNN, Logistic Regression, and SVM, for fake review detection.
- **To show** the comparison of the models' performance based on metrics like accuracy, precision, recall, and F1-score, identifying the best-performing algorithm for detecting fake reviews.

1.4 Scope

This venture will zero in on identifying counterfeit inn surveys utilizing an AI based approach. It will include preprocessing a lodging surveys dataset, carrying out and looking at various calculations, and assessing their presentation with regards to order exactness. The undertaking will be restricted to the utilization of NLTK and scikit-learn libraries for text handling and model execution. The last model will be pointed toward helping purchasers in going with informed choices by sifting through counterfeit audits and assisting lodging the board with keeping up with the validity of their web-based standing.

This project is dedicated to the identification of fraudulent hotel reviews through the application of machine learning algorithms. It will encompass data preprocessing, feature extraction, and the deployment of models including Decision Tree, KNN, Logistic Regression, and SVM. The focus is restricted to utilizing the NLTK and scikit-learn libraries for text analysis and model construction. The objective is to determine the most effective algorithm for categorizing hotel reviews as either authentic or fraudulent. Furthermore, the project will investigate optimization strategies, such as hyperparameter tuning and cross-validation, to improve detection precision. The ultimate goal of the system is to aid consumers in making better-informed choices by eliminating fake reviews, while also assisting hotel managers in maintaining the integrity of their online reputation.

In addition to these primary tasks, the project may provide insights into how various features within hotel reviews—such as sentiment analysis, review trends, and user behavior—affect the efficacy of fake review detection. This research could prove beneficial for enhancing future models and tackling the increasing prevalence of deceptive online reviews in the hospitality sector.

Chapter 2

Review of Literature

Sr. No.	Title of paper	Database used	Algorithms & methods used	Research Gap
2.1	An Empirical Study on Detecting Fake Reviews Using Machine Learning Techniques[3]	Movie reviews dataset	Sentiment Analysis, NB, DT-J48, KNN-IBK and SVM	Collusion and manipulation issues weren’t addressed.
2.2	Spotting and Removing Fake Product Review in Consumer Rating Reviews[2]	online shopping reviews dataset	Support Vector Machine (SVM)	Comparison of various algorithms.
2.3	Sentiment analysis on Chinese movie review with distributed keyword vector representation[4,5]	Chinese movie review dataset	TF-IDF and LLR NB, DT-J48, KNN-IBK and SVM	Accuracy is not quite good
2.4	Fake Review Detection on Yelp Dataset Using Classification Techniques in Machine Learning[7]	Dataset of reviews on restaurants and hotels	Logistic Regression, NB, XGBoost, SVM	Imbalance in the dataset is to be handled.
2.5	Fake review detection review using Behavioural and Contextual features[5]	pseudo fake ,real life reviews	SVM,decision trees ,random forest	Limited dataset
2.6	A Method for the Detection of Fake Reviews Based on Temporal Features of Reviews and Comments [8]	Amazon-China shopping dataset	Isolation forest, bayesian network	Product recommendation

Table 2.1 Research papers that were studied.

Chapter 3

System Architecture

3.1. Block Diagram:

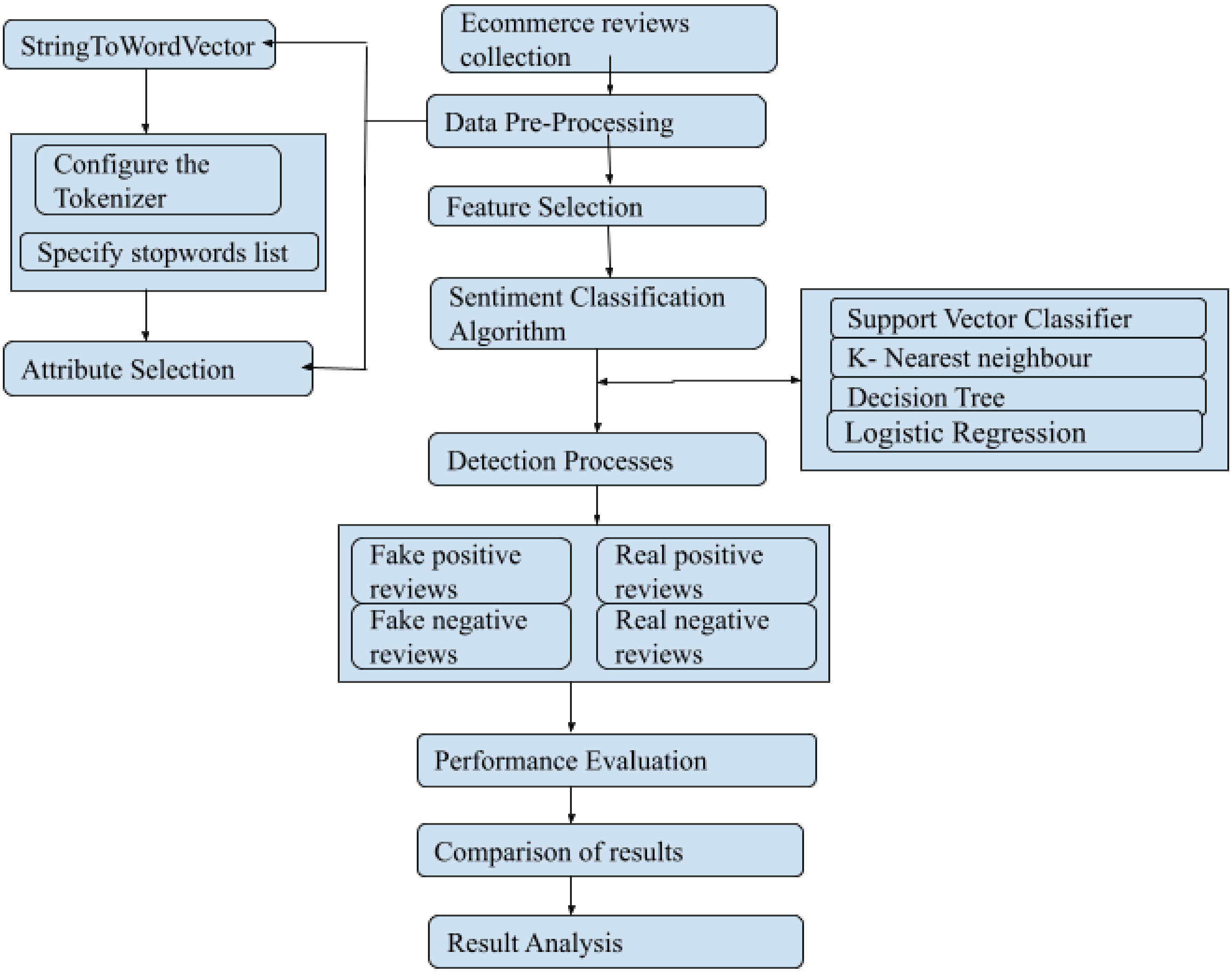


Figure. 3.1 Block Diagram

3.2. SYSTEM DESIGN:

3.2.1 Use Case Diagram

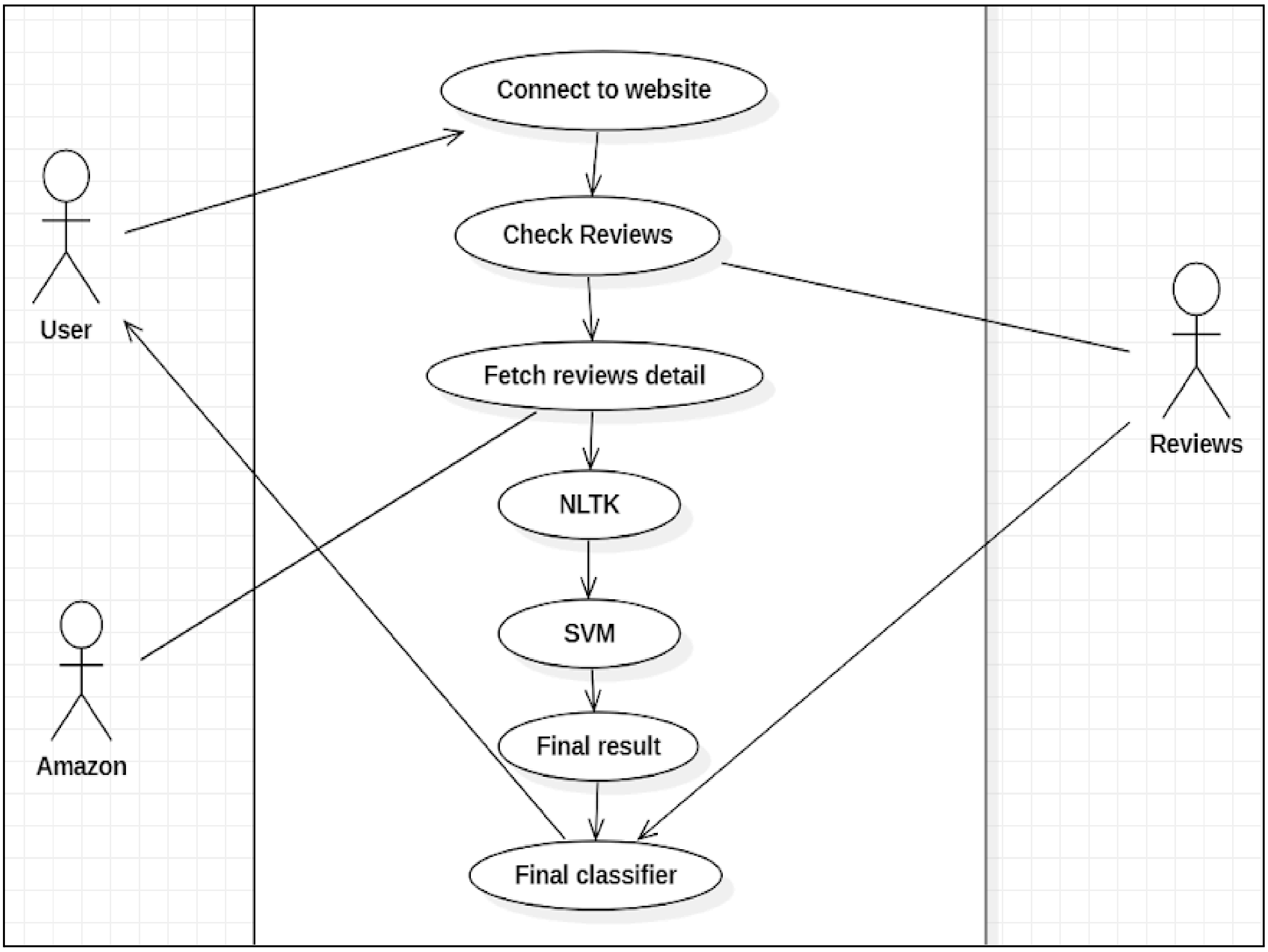


Figure. 3.2. Use Case Diagram

3.2.2 DATA FLOW DIAGRAM(LEVEL 0)

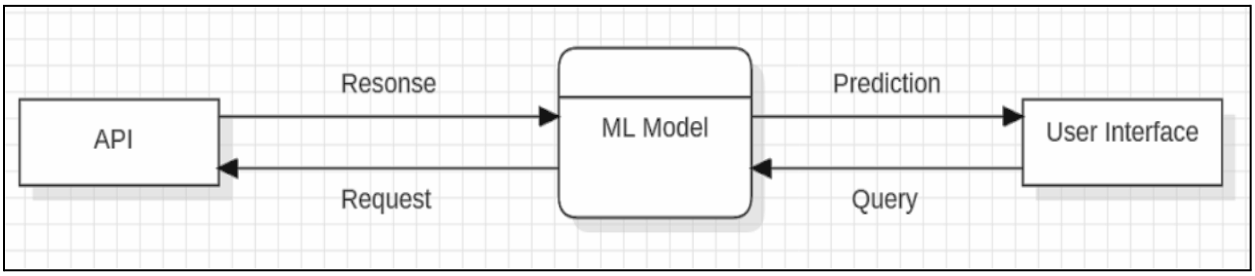


Figure. 3.3. Data Flow Diagram (Level 0)

3.2.3 DATA FLOW DIAGRAM(LEVEL 1)

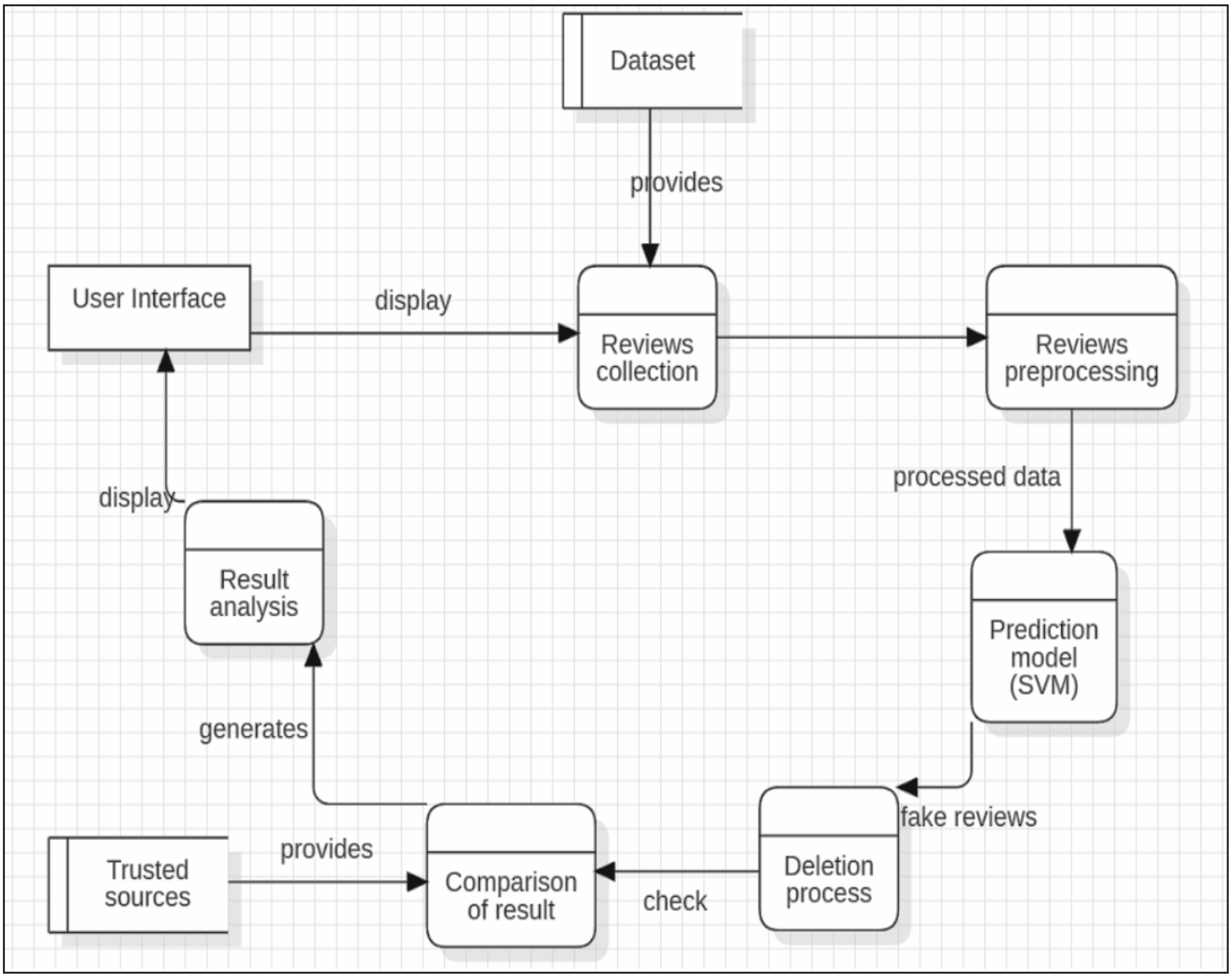


Figure. 3.4. Data Flow Diagram (Level 1)

In 1-level DFD, a context diagram is decomposed into multiple bubbles/processes. In this level, we highlight the main objectives of the system and break down the high-level process of 0-level DFD into subprocesses.

In the Level 1 Data Flow Diagram (DFD) for the project focused on detecting fraudulent hotel reviews, the primary processes encompass the reception of the hotel review dataset as input, followed by a preprocessing stage where the data undergoes cleaning and the extraction of pertinent features. The refined data is subsequently utilized to train various machine learning models, including Decision Tree, K-Nearest Neighbors (KNN), Logistic Regression, and Support Vector Machine (SVM). After the training phase, the models are fine-tuned to enhance their performance. These optimized models are then employed to categorize new reviews as either fraudulent or authentic. Ultimately, the outcomes are assessed, and performance metrics are produced, offering valuable insights into the system's accuracy and efficacy in identifying fake reviews. The external entities involved in this process include the input dataset and the output consisting of classified reviews along with performance metrics.

3.2.4 DATA FLOW DIAGRAM (LEVEL 2)

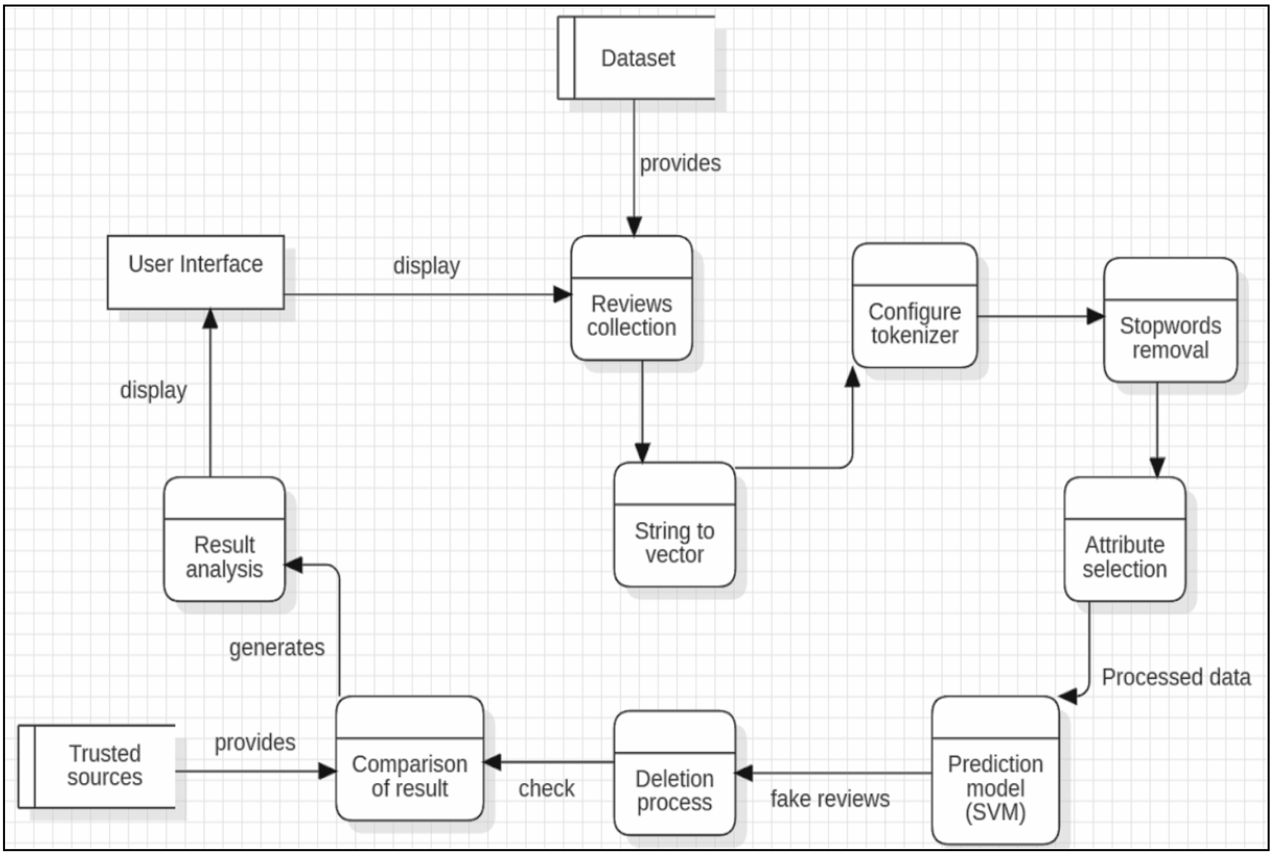


Figure. 3.5. Data Flow Diagram (Level 2)

2-level DFD goes one process deeper into parts of 1-level DFD. It can be used to project or record the specific/necessary detail about the system's functioning.

In the Level 2 Data Flow Diagram (DFD) for the project focused on detecting fraudulent hotel reviews, the initial step involves the input of the hotel review dataset. This dataset undergoes a series of preprocessing steps, which encompass addressing missing values, text cleaning, and feature extraction utilizing the Natural Language Toolkit (NLTK). Subsequently, the refined dataset is employed to train a variety of machine learning algorithms, including Decision Tree, K-Nearest Neighbors (KNN), Logistic Regression, and Support Vector Machine (SVM). Following the training phase, the models are fine-tuned through hyperparameter optimization and cross-validation to enhance their performance. The optimized models are then applied to identify fake reviews, categorizing new submissions as either fraudulent or authentic. Ultimately, the outcomes, which include performance metrics such as accuracy and F1-score, are assessed and documented, providing insights into the system's effectiveness.

3.2. Dataset:

It is a corpus of truthful and deceptive hotel reviews of 20 Chicago hotels. The data is described in two papers according to the sentiment of the review. In particular, positive sentiment reviews in [9] and negative sentiment reviews in [3]. The dataset is provided by cornell.edu.in.

This corpus contains:

- 400 truthful positive reviews from TripAdvisor (described in [2])
- 400 deceptive positive reviews from Mechanical Turk (described in [2])
- 400 truthful negative reviews from Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor and Yelp (described in [3])
- 400 deceptive negative reviews from Mechanical Turk (described in [3])

Each of the above data sets consist of 20 reviews for each of the 20 most popular Chicago hotels (see [2] for more details).

Hotels included in this dataset:

- affinia: Affinia Chicago (now MileNorth, A Chicago Hotel)
- allegro: Hotel Allegro Chicago - a Kimpton Hotel
- amalfi: Amalfi Hotel Chicago
- ambassador: Ambassador East Hotel (now PUBLIC Chicago)
- conrad: Conrad Chicago
- fairmont: Fairmont Chicago Millennium Park
- hardrock: Hard Rock Hotel Chicago
- hilton: Hilton Chicago
- homewood: Homewood Suites by Hilton Chicago Downtown
- hyatt: Hyatt Regency Chicago
- intercontinental: InterContinental Chicago
- james: James Chicago

- knickerbocker: Millennium Knickerbocker Hotel Chicago
- monaco: Hotel Monaco Chicago - a Kimpton Hotel
- omni: Omni Chicago Hotel
- palmer: The Palmer House Hilton
- sheraton: Sheraton Chicago Hotel and Towers
- sofitel: Sofitel Chicago Water Tower
- swissotel: Swissotel Chicago
- talbott: The Talbott Hotel

Chapter 4

System Implementation

4.1. Requirements:

4.1.1 Software Requirements:

- Visual Studio Code/Jupyter Notebook/Google Collabs/PyCharm for executing programs
- Excel for Database.

4.1.2 Hardware Requirements:

- OS: Windows 7 and above
- RAM: 8GB and above (Preferable 32GB), Hard-disk: 250GB and Above
- GPU: 2GB and above
- Network: Internet Connectivity

4.1.3 Technologies Used:

- Python
- Jupyter Notebook
- Machine Learning, Sentiment Analysis

4.2. Algorithms:

4.2.1 Decision Tree :

Decision Tree solves the problem of machine learning by transforming the data into a tree representation. Each internal node of the tree representation denotes an attribute and each leaf node denotes a class label.A decision tree algorithm can be used to solve both regression and classification problems. The Accuracy of Algorithm for our dataset comes out to be 0.6475.

The Decision Tree algorithm has the following advantages:

- Compared to other algorithms, decision trees require less effort for data preparation during pre-processing.
- A decision tree does not require normalization of data.
- A decision tree does not require scaling of data as well.
- Missing values in the data also do NOT affect the process of building a decision tree to any considerable extent.
- A Decision tree model is very intuitive and easy to explain to technical teams as well as stakeholders.

The Decision Tree algorithm has the following limitations:

- A small change in the data can cause a large change in the structure of the decision tree causing instability.
- For a Decision tree sometimes calculation can go far more complex compared to other algorithms.
- Decision trees often involve higher time to train the model.
- Decision tree training is relatively expensive as the complexity and time taken are more.
- The Decision Tree algorithm is inadequate for applying regression and predicting continuous values.

4.2.2 Logistic Regression:

Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a data set. . A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. The Accuracy of Algorithm for our dataset comes out to be 0.90.

$$E(w,b) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)) + \alpha R(w)$$

where, L is a loss function that measures model (mis)fit and R is a regularization term (aka penalty) that penalizes model complexity; α>0 is a non-negative hyperparameter that controls the regularization strength.

The Logistic Regression algorithm has the following advantages:

- Logistic Regression performs well when the dataset is linearly separable.
- Logistic regression is less prone to overfitting but it can overfit in high dimensional datasets. You should consider Regularization (L1 and L2) techniques to avoid over-fitting in these scenarios.[15]
- Logistic Regression not only gives a measure of how relevant a predictor (coefficient size) is, but also its direction of association (positive or negative).[13]
- Logistic regression is easier to implement, interpret and very efficient to train.

The Logistic Regression algorithm has the following limitations :

- Main limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables. In the real world, the data is rarely linearly separable. Most of the time data would be a jumbled mess.
- If the number of observations are lesser than the number of features, Logistic Regression should not be used, otherwise it may lead to overfit.
- Logistic Regression can only be used to predict discrete functions. Therefore, the dependent variable of Logistic Regression is restricted to the discrete number set. This restriction itself is problematic, as it is prohibitive to the prediction of continuous data.[1]

4.2.3 K-Nearest Neighbour :

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. The K-NN algorithm assumes the similarity between the new case/data and available cases and puts the new case into the category that is most similar to the available categories.The Accuracy of Algorithm for our dataset comes out to be 0.790.

The K-NN working can be explained on the basis of the below algorithm:

- **Step-1:** Select the number K of the neighbors
- **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
- **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
- **Step-4:** Among these k neighbors, count the number of the data points in each category.
- **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.
- **Step-6:** Our model is ready.

4.2.4 Support Vector Machine:

Support vector machines or SVM is a supervised machine learning algorithm that can be used for both classification and regression analysis. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes.[13][9] The Accuracy of Algorithm for our dataset comes out to be 0.905.

The RBF kernel function for two points X_1 and X_2 computes the similarity or how close they are to each other. This kernel can be mathematically represented as follows:

$$K(X_1, X_2) = exp(-\frac{||X_1 - X_2||^2}{2\sigma^2})$$

where,

1. ‘σ’ is the variance and our hyperparameter
2. $||X_1 - X_2||$ is the Euclidean (L_2-norm) Distance between two points X_1 and X_2

The SVM algorithm has the following advantages:

- SVM's are very good when we have no idea on the data.
- Works well with even unstructured and semi structured data like text, Images and trees.[6]
- The kernel trick is the real strength of SVM. With an appropriate kernel function, we can solve any complex problem.
- Unlike in neural networks, SVM is not solved for local optima.
- It scales relatively well to high dimensional data.[5]
- SVM models have generalization in practice, the risk of overfitting is less in SVM.
- SVM is always compared with ANN. When compared to ANN models, SVMs give better results.[14]

The SVM algorithm has the following limitations :

- Choosing a "good" kernel function is not easy.
- Long training time for large datasets.
- Difficult to understand and interpret the final model, variable weights and individual impact.[9]
- Since the final model is not so easy to see, we can not do small calibrations to the model hence its tough to incorporate our business logic.
- The SVM hyperparameters are Cost -C and gamma. It is not that easy to fine-tune these hyper-parameters. It is hard to visualize their impact.

4.3. Parameters Estimated:

- **Accuracy** is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. For our model, we have got 0.905 which means our model is approx. 80% accurate.[5]

Accuracy = TP+TN/TP+FP+FN+TN

- **Precision** - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. High precision relates to the low false positive rate. We have got 0.903 precision which is pretty good.

Precision = TP/TP+FP

- **Recall** (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class. We have got a recall of 0.631 which is good for this model as it's above 0.912. [12][8]

Recall = TP/TP+FN

- **F1 score** - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.[3] Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall. In our case, the F1 score is 0.907. [11]

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

4.4. Modules Implemented:

A) NumPy: NumPy is a well known general-purpose array-processing package. An extensive collection of high complexity mathematical functions make NumPy powerful enough to process large multi-dimensional arrays and matrices.With NumPy, you can define arbitrary data types and easily integrate with most databases. NumPy can also serve as an efficient multi-dimensional container for any generic data that is in any datatype.[5]

B) Scikit-learn: Scikit-learn has a wide range of supervised and unsupervised learning algorithms that works on a consistent interface in Python. The library can also be used for data-mining and data analysis. The main machine learning functions that the Scikit-learn library can handle are classification, regression, clustering, dimensionality reduction, model selection, and preprocessing.[6][9]

C) **Pandas :**Pandas are turning up to be the most popular Python library that is used for data analysis with support for fast, flexible, and expressive data structures designed to work on both “relational” or “labeled” data. Pandas today is an inevitable library for solving practical, real-world data analysis in Python. Pandas is highly stable, providing highly optimized performance. [3][1]

D) **Matplotlib:** Matplotlib is a data visualization library that is used for 2D plotting to produce publication-quality image plots and figures in a variety of formats. The library helps to generate histograms, plots, error charts, scatter plots, bar charts with just a few lines of code.It works by using standard GUI toolkits like GTK+, wxPython, Tkinter, or Qt to provide an object-oriented API that helps programmers to embed graphs and plots into their applications.

Chapter 5

Results and Discussion

Sr. No.	Algorithm Implemented	Accuracy	Precision	Recall	F1 Score
1.	Support Vector Classifier	0.905	0.903	0.912	0.908
2.	K-Nearest Neighbors	0.790	0.885	0.678	0.768
3.	Decision Tree	0.647	0.693	0.561	0.619
4.	Logistic Regression	0.900	0.895	0.912	0.903

Table 5.1 Comparison of different algorithms

I have discussed different fake reviews detection techniques that are based on supervised learning methodologies. Table 5.1 shows the results arrived at by calculating the accuracies of the various models mentioned above. The accuracy, precisions and f1 score can be computed with the help of confusion matrices. A single confusion matrix was created for each model.

I have compared 4 different techniques, (Support Vector Classifier, K-Nearest Neighbors, Decision Tree, Logistic Regression), to identify fake reviews. of which SVM has proved to be the best with accuracy of 0.905.

The values shown are the averaged values over successive trials. Based on the results in Table 3.1, the graph in Fig. 3.1 is constructed by taking different algorithms on X-axis and accuracy on the Y-axis. It is inferred that SVM provides us with the highest accuracy followed by LR

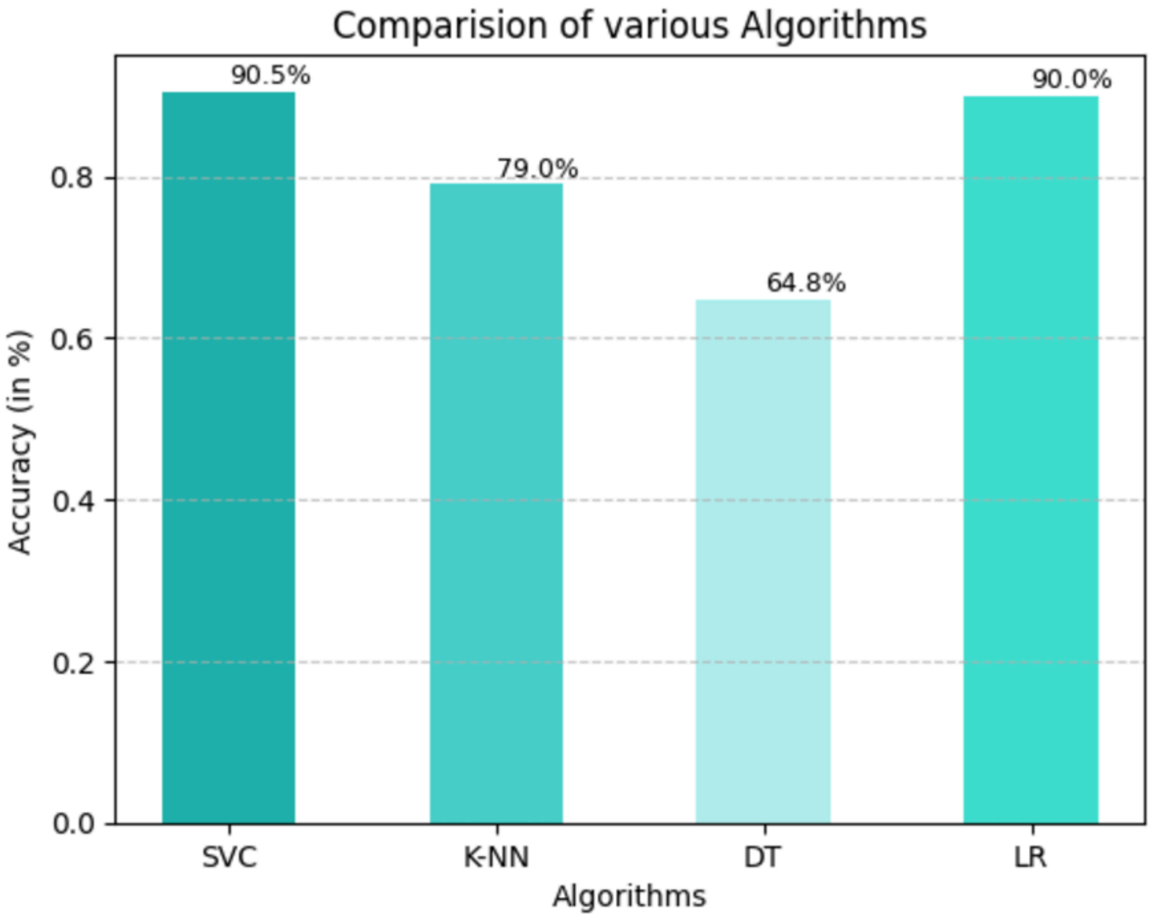
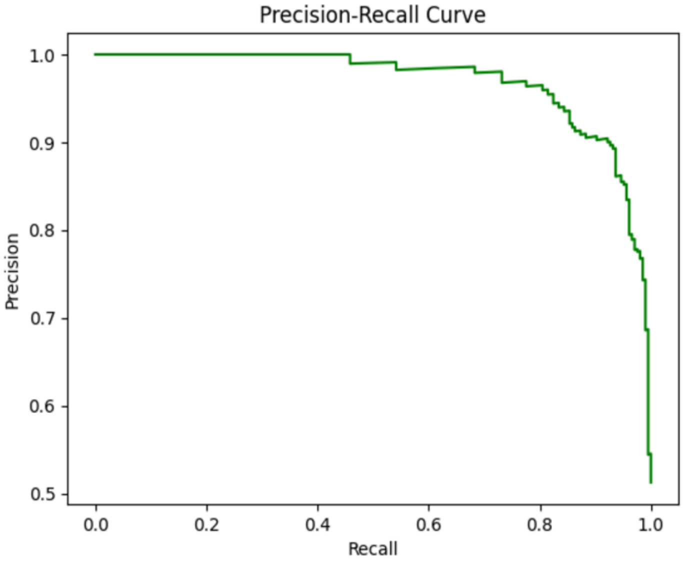


Figure. 5.1. Graph of different algorithms



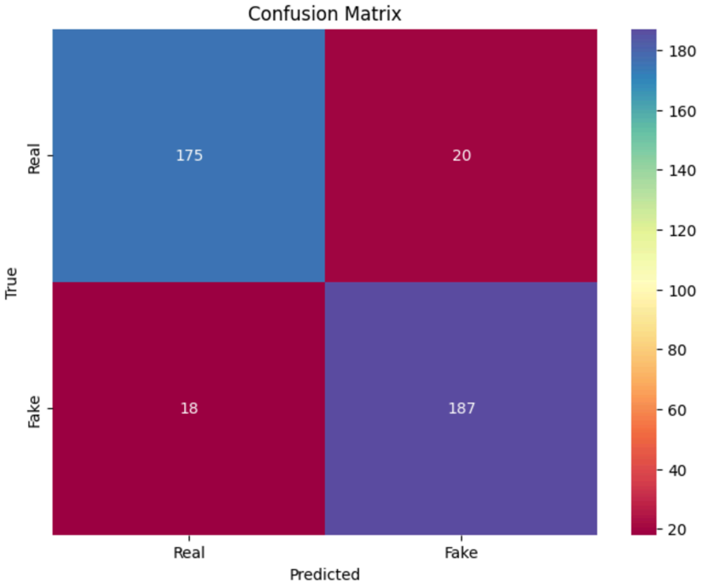


Figure. 5.2. Precision Recall Curve

Figure. 5.3. Confusion Matrix

Chapter 6

Conclusion And Future Scope

Due to rapid development of the internet, the size of the reviews of the items / products increases. These huge amounts of information are generated on the Internet; there is no analysis of the quality of reviews that are written by consumers. Anyone can write anything which conclusively leads to fake reviews or some companies are hiring people to post reviews. Some of the fake reviews that have been intentionally fabricated to seem genuine, capability to identify fake online reviews are crucial.

In this paper, we have presented a model for fake review detection through different machine learning techniques. Furthermore, the paper investigated the seven methods and compared their accuracy and four basic kernels. The model that achieves the highest accuracy is SVM using the RBF kernel and the highest accuracy score is 90.5%.

Fake review detection is an emerging research area that has a scarce number of datasets. There is no data on real-time news or regarding current affairs. The current model is trained using the existing dataset which shows that the model performs well against it.

Utilizing a moderately bigger dataset to prepare the framework can be one of the things to come for our venture. The subsequent stage then, at that point, is to train the model and investigate how the accuracy changes with the new dataset to develop it further.

References

[1] Parikh, S. B., & Atrey, P. K. (2018, April).Media-Rich Fake News Detection: A Survey. In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 436-441). IEEE.

[2] Y. Li, X. Feng, and S. Zhang, “Detecting Fake Reviews Utilizing Semantic and Emotion Model,” 2016 3rd International Conference on Information Science and Control Engineering (ICISCE), pp. 317–320, 2016.

[3] J. Kamps, M. Marx, R.J. Mokken, and M. Rijke, “Using WordNet to measure semantic orientations of adjectives,” Proceedings of the Fourth International Conference on Language Resources and Evaluation, vol. IV, pp 1115-1118, 2004.

[4] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment classification using machine learning techniques,” Proceedings of EMNLP, pp. 79-86, 2002.

[5] C. Castillo, M. Mendoza, and B. Poblete, “Predicting information credibility in time-sensitive social media,” Internet Research, vol. 23, no. 5, pp. 560–588, 2012.

[6] M. Singh, L. Kumar, and S. Sinha, “Model for Detecting Fake or Spam Reviews,” Advances in Intelligent Systems and Computing ICT Based Innovations, pp. 213–217, Jan. 2017.

[7] N. Jindal and B. Liu, "Analyzing and Detecting Review Spam," Seventh IEEE International Conference on Data Mining (ICDM 2007), 2007, pp. 547-552, doi: 10.1109/ICDM.2007.68.

[8] Wang, JZ, Z Yan, LT Yang and BX Huang, “An approach to rank reviews by fusing and mining opinions based on review pertinence”, pp.3–15, 2015.

[9] T. J. Ma and D. Atkin, “User-generated content and credibility evaluation of online health information: A meta-analytic study,” Telematics and Informatics, 2016.

[10] M. Hu and B. Liu, “Mining and summarizing customer reviews,” Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 168-177, 2004.

[11] B. J. Fogg and H. Tseng, “The elements of computer credibility,” in Proc. of the SIGCHI Conf. on Human Factors in Computing Systems. ACM, 1999, pp. 80–87.

[12] C. Castillo, M. Mendoza, and B. Poblete, “Predicting information credibility in time-sensitive social media,” Internet Research, vol. 23, no. 5, pp. 560–588, 2012.

[13] T. J. Ma and D. Atkin, “User generated content and credibility evaluation of online health information: A meta analytic study,” Telematics and Informatics, 2016.

[14] Mukherjee, V. Venkataraman, B. Liu, and N. S. Glance, “What Yelp fake review filter might be doing?” in Proceedings of ICWSM, 2013.

[15] A. Parikh, S. B., & Atrey, P. K. (2018, April).Media-Rich Fake News Detection: A Survey. In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 436-441). IEEE.

Acknowledgment

I would like to take this opportunity to thank everyone who has helped me along the way and whose contributions have made this project report possible. I want to start by expressing my sincere gratitude for my project guide, **Ms. Shraddha More** (**Deputy HOD, Associate Professor**, Department of Information Technology, St. John College of Engineering and Management, Palghar), for her unwavering support, encouragement, and helpful criticism and comments throughout the discussions, all of which greatly aided me in finishing this project.

Additionally, I am grateful to **Dr. Nilesh T. Deotale, Head of Department** at St. John College of Engineering and Management in Palghar, who has consistently provided the best support and assistance.

The completion of this project report would not have been possible without the assistance of **Dr. Kamal Shah, Principal** of St. John College of Engineering and Management, Palghar, and all of the staff members.