

# Fake Profile Detection on Matrimonial Sites using Random Forest Classifier

<sup>[1]</sup> Prof. Jagruti Nagoankar, <sup>[2]</sup> Riddhi Dinesh Oza, <sup>[3]</sup> Ayukshi Rane, <sup>[4]</sup> Godsee Almeida, <sup>[5]</sup> Shaun Dabre

<sup>[1]</sup> <sup>[2]</sup> <sup>[3]</sup> <sup>[4]</sup> <sup>[5]</sup> Department of Computer Engineering, Fr. Conceicao Rodrigues College of Engineering,  
Mumbai, Maharashtra, India

Corresponding Author Email: <sup>[1]</sup> jagruti@frcrce.ac.in, <sup>[2]</sup> riddhioza1779@gmail.com, <sup>[3]</sup> ayukshirane18@gmail.com,  
<sup>[4]</sup> godseealmeida@gmail.com, <sup>[5]</sup> shaun.dabre9291@gmail.com

**Abstract—** Due to immense growth in Matrimonial Industry, people from various religions, caste, educational backgrounds, economic status and varied regions got a platform to look for someone with similar interests. Though this evolution has led to many successful marriages, at the same time there has been intervention of some unwanted profiles into the matrimonial sites. People who don't really want to marry but for the sake of monetary benefits, they make fake profiles and try to fool people with genuine intentions. This also causes threat to the reputation of the organization. Thus, there is a need to take into account such profiles and create awareness regarding the same.

**Index Terms—** matrimonial site, fake profiles, detection of fake profiles.

## I. INTRODUCTION

Evolution of internet has significantly changed each and every sector, thus giving new opportunities to business in various possible ways; from banking, railway reservations, hospitals and many more. One such industry which has been impacted due to the adoption of internet is the Matrimony industry. Before internet was not so common among the people, traditional marriage bureaus, personal meetings within families, relations with different people etc. were some of the ways in which marriages were carried forward, providing limited options to the person who wished to marry. As the use of internet increased and knocked the doors of more and more people, marriage bureaus converted to "matrimonial sites", thus giving ample of choices to the people who desired to marry. On matrimonial sites, though there are many users who have genuine intentions to find a match, there are some people with bad intentions as well who try to misuse such platforms for monetary benefits. To solve this issue, we kept in mind all the possible aspects that a fake profile can display. We took the help of some of the renowned matrimonial sites who have large customer base with many successful stories.

## II. LITERATURE REVIEW

Due to the bloom in the use of internet, demand for online matrimonial sites increased significantly. This can be well understood by "A Study on Digitalization of Matchmaking for Marriages in India"[1], where the study focuses on how internet revolutionised the matrimony industry and the ways in which Matrimonial sites make use of techniques such as SEO(Search Engine Optimization) & SEM(Search Engine Marketing) for acquiring instant leads. A survey of 100 brides and grooms, who have registered on matrimonial

portals was conducted to study why online matrimonial sites are preferred over traditional match making process in "Study on the factors affecting the preferences of online matrimonial portals over traditional methods"[2], which concluded that search, convenience, website features and trust are some of the important factors influencing the customer satisfaction of the user. Though online matrimonial portals are convenient but it attracts many people with bad intentions thus having adverse effects on reputation of such organizations. This problem also persists on social networks like LinkedIn. Approach proposed by "Identifying fake profiles in LinkedIn"[3], can identify fake profiles with an accuracy of 87% using SVM with Polynomial Kernel using PCA-selected features. In order to identify legitimate users on matrimonial sites, "Towards Safe Spaces Online: A Study of Indian Matrimonial Websites"[4], focuses on the authentication of users under the law established by the government of India. Access will not be granted until the user authenticates. The profile creator can be verified by providing an email id or phone number, and a verification token (a link or OTP) will be sent to that phone number or email id. The purpose of this step is to verify that the email address or phone number belongs to the profile creator who is requesting an account on the portal. In the thesis proposed in "Detecting Fake Profiles on Online Matrimony"[5], focus is given on finding out the unique characteristics of a fake profile. For detection autoencoders have been used which outperforms the feature engineering methods. 5 The article "Fake Profile Detection on Social Networking Websites: A Comprehensive Review[6]," aims to summarize the recent advancement in the fake account detection methodology on social networking websites. In the past few years, researchers applied many advanced technologies to identify fake accounts. In the survey presented in this article, the recent development of fake account detection technologies has been

summarized. We discuss the challenges and limitations of the existing models in brief. The survey may help future researchers to identify the gaps in the current literature and develop a generalized framework for fake profile detection on social networking. The paper "Social Networks Fake Profiles Detection Using Machine Learning Algorithms"[7], sheds light on the role of fake identities in advanced persistent threats and covers the mentioned approaches of detecting fake social media profiles. In order to make a relevant prediction of fake or genuine profiles, the impact of three supervised machine learning algorithms: Random Forest (RF), Decision Tree (DT-J48), and Naïve Bayes (NB) have been assessed. In the paper "A Hybrid Method for Fake Profile Detection in Social Network Using Artificial Intelligence"[8], techniques for Artificial Intelligence and Natural Language Processing (NLP) techniques to increase accuracy of fake profile recognition are suggested. We use the Random Forest Classifier, Support Vector Machine (SVM) and Optimized Naïve Bayes algorithm to categorize profiles into fake or genuine classes. These three algorithms were used to determine the true or false identity of the target accounts. This algorithm uses less features, but about 98% of training dataset accounts can still be correctly defined. "Detection of fake profiles using machine learning"[9], focuses not only on fake profile detection but also on the execution time while determining fake profiles. It concludes that similarity-based approach is less time consuming and is more efficient for detection of fake profiles. In approach proposed in "Financial Fraud Detection Model: Based on Random Forest"[12], random forest proves to provide the highest accuracy along with the use of multi-dimensional scaling. This idea is also supported in "Detection of Fake Profiles on Online Matrimony Using Machine Learning and Deep Learning"[11]. In article, "Commercial Utilization of Chat-Bot Using Python"[12], a detailed study on use of chat bots have been made. A chat-bot is a computer programmed based on the AIML (artificial intelligence markup) and NLP. Any questions will be asked by an individual and chat-bot will reply. A chat-bot is now extremely common and uses speed as a communication application for computers. For any business website the Chat-bot framework is in demand, thus it is used on numerous websites. The main aim of this chat-bot is to build a web API, sample web and text messaging interfaces that illustrates the use of a GUI.

### III. PROBLEMS WITH THE EXISTING SYSTEMS

To identify the problems with the existing systems and come up with the appropriate solution, we surveyed India's three leading matrimonial sites: Shaadi.com, Jeevansathi.com and Bharat Matrimony.

The problem with the existing matrimonial sites as per real life experience is:

- 1) Easy access to online matrimonial portals attracts many people with bad intentions. This ends up in the creation of lots of fake profiles.

- 2) Among all the existing matrimonial sites, very few of them have facilities for document verification and validating the user.
- 3) Matrimonial sites' recommendation system keeps on flooding the profiles in the inbox of the user even though they don't match with the criteria of the user.
- 4) Almost all the sites ask for excessive details in one go, thus it ends up losing the customer base.

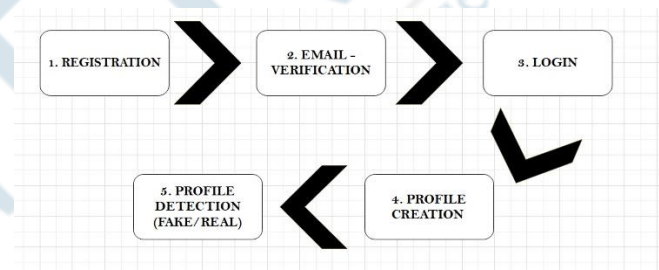
### IV. PROPOSED METHODOLOGY

The module diagram of the entire system is shown in (Fig. 1).

1). Following are the modules used in our project:

- 1) Registration
- 2) Email-Verification
- 3) Login
- 4) Profile Creation
- 5) Profile Detection (fake/real)

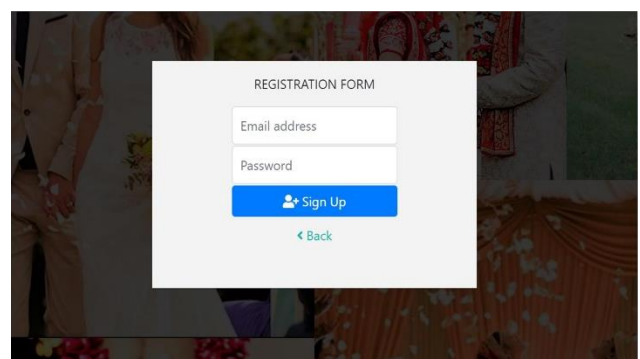
The sections that follow give in depth descriptions of each of the modules used in the system.



**Fig. 1: Module Diagram**

#### A. Registration

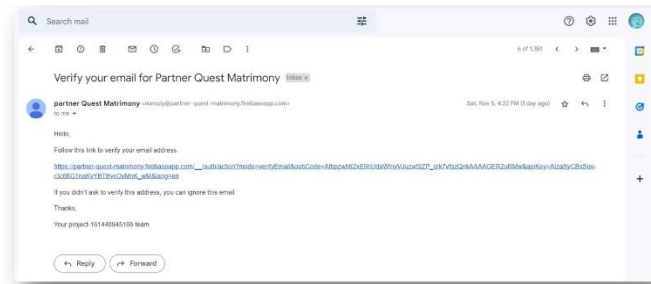
For creating an account on Partner Quest (matrimonial site), the user needs to register by filling out appropriate details (an email ID and a password of at least six characters) in the registration form. The registration page is shown in (Fig. 2).



**Fig. 2: Registration Page**

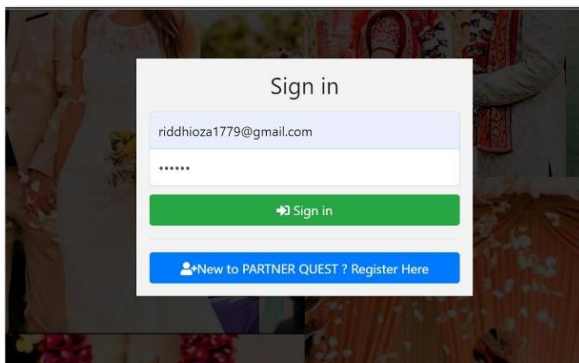
#### B. Email-Verification

After filling out the form and clicking on the "sign up" button, a verification email would be sent, and upon clicking that link, the user would be verified. The demonstration of email-verification is shown in (Fig. 3).

**Fig. 2: Email-Verification**

### C. Login

For login, the user needs to enter their email address and password and click on the "sign in" button. For login and registration with authentication, Firebase is used. User details would get stored in the database using MySQL's phpMyAdmin. The Login Page is shown in (Fig. 4).



**Fig. 3: Login Page**

### D. Profile Creation

The user who has successfully registered and logged in to the Partner Quest website, need to create his/her profile. The website asks the user to fill the following details:

- 1) Caste
- 2) Degree
- 3) Employment Sector
- 4) Income
- 5) Mother Tongue
- 6) Occupation
- 7) Religion

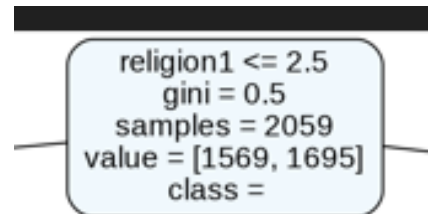
### E. Profile Detection (Fake/Real)

Once the user has filled in the required details for creating profile, he/she will be classified as fake/real depending upon the inputs provided.

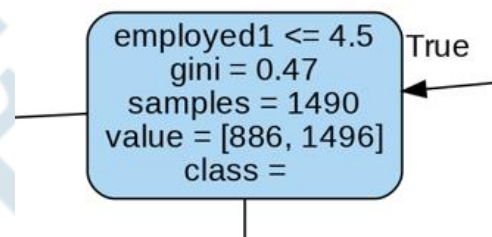
### F. Working of the entire system

The model is provided with the data sets of both real and fake profiles. The data set has been collected from [13] using web scraping and originally consisted of 67 columns, using the concepts of dimensionality reduction and feature selection we have considered eight most significant columns corresponding to the seven attributes: “caste”, “degree”, “employment sector”, “income”, “mother tongue”,

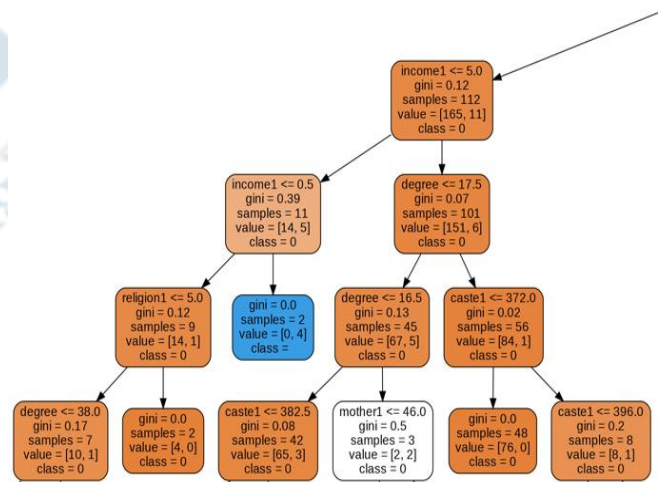
“occupation”, “religion” and “account (fake/real)”. Once the model is trained with the given data set, it is given the new inputs by the “profile creation” page. Through the inputs provided by the user, model tries to predict whether the profile might be fake or real. The following figures show the decision trees generated by our model.



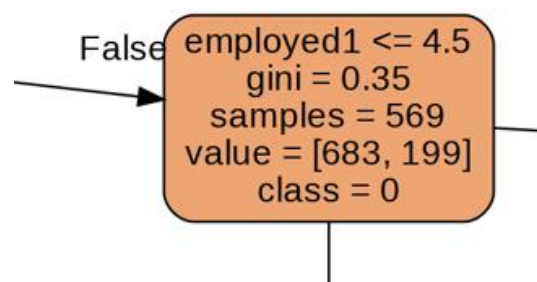
**Fig. 4:** Root Node of Decision Tree



**Fig. 5: Node when Root Node calculations results in True**



**Fig. 6:** Part of Left Sub-Tree of the tree generated



**Fig. 7:** Node when Root Node calculations result in False



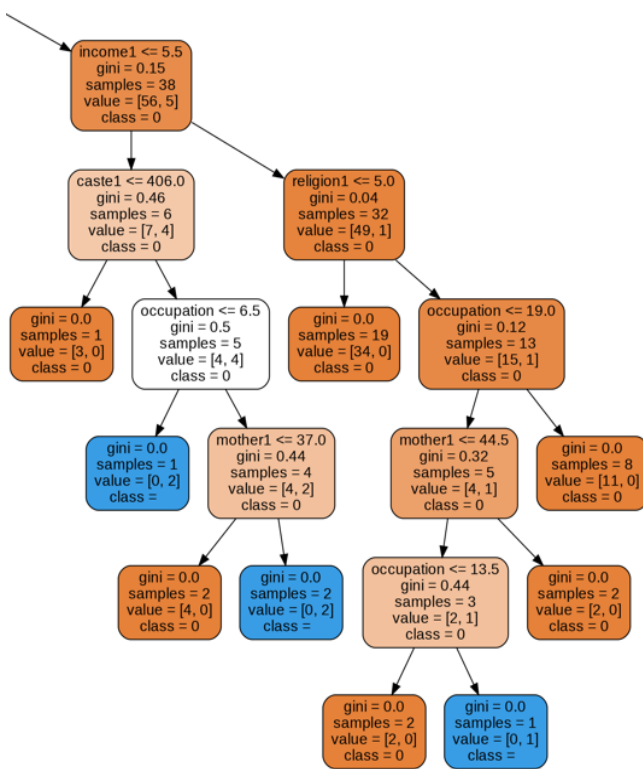


Fig. 8: Part of Right sub-tree of the tree generated

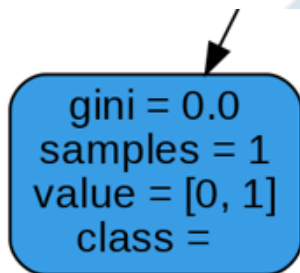


Fig. 9: class= indicates Real Profile

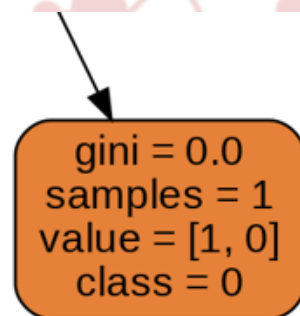


Fig. 10: class=0 indicates Fake Profile

```
[ ] print(f'Model Accuracy: {rfc.score(X, y)}')
```

Model Accuracy: 0.978448275862069

Fig. 11: Model Accuracy using Random Forest

```
[ ] print(f'Model Accuracy: {svclassifier.score(X, y)}')
```

Model Accuracy: 0.6570197044334976

Fig. 12: Model Accuracy using Support Vector Machine (SVM)

```
from statistics import mean
mean(accuracy_ann)
```

0.8655312657356262

Fig. 13: Module Accuracy using Artificial Neural Networks (ANN)

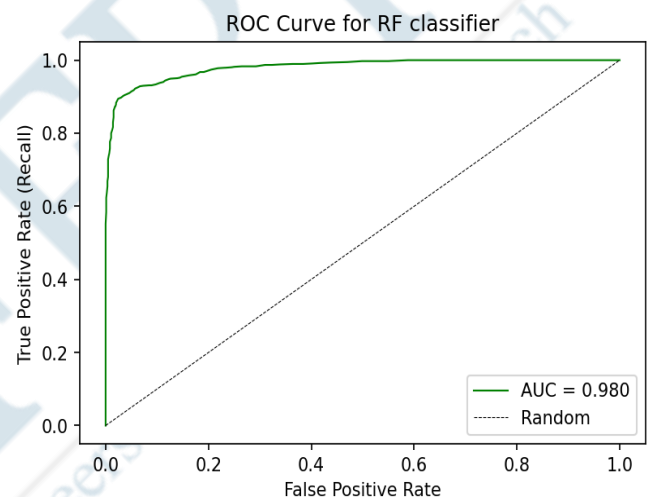


Fig. 14: Receiver Operating Characteristic (ROC) for Random Forest

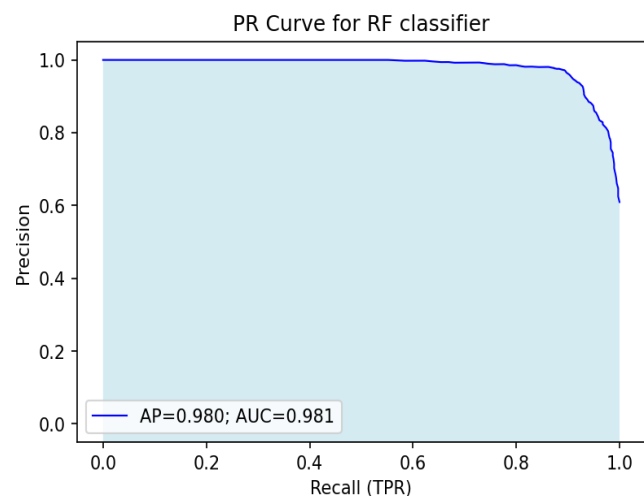


Fig. 15: Precision of the model for Random Forest

## V. PERFORMANCE EVALUATION

### G. Algorithms Used

**Table 1:** Description of Algorithms Used

| Parameter                    | SVM<br>(Support Vector Machine)                       | NN<br>(Neural Networks)                                                   | Random Forest Classifier                                         |
|------------------------------|-------------------------------------------------------|---------------------------------------------------------------------------|------------------------------------------------------------------|
| Input Data                   | Performs well with small amount of data.              | Performs well with large amount of data                                   | Performs well with small amount of data.                         |
| Pre-processing of Input Data | Requires minimal/less processing, thus saves time     | Requires lots of data cleaning and modification which is time consuming   | Requires minimal/less processing, thus saves time.               |
| Total number parameters      | Machine considers following two parameters: C & gamma | Depends on: Number of Hidden layers, Learning Rate & iterations or epochs | Depends on: Depth of the tree & No. Of Tress/Nodes in each layer |
| Accuracy                     | 0.6570197044334976                                    | 0.8655312657356262                                                        | 0.9782430213464697                                               |

### H. Justification for the chosen Algorithm

**Table 2:** Algorithms Used

| Algorithms Used                            |                                       |                                       |
|--------------------------------------------|---------------------------------------|---------------------------------------|
| 1. Support Vector Machine                  | 2. Artificial Neural Networks         | 3. Random Forest Classifier           |
| Supervised Machine Learning Algorithm      | Supervised Machine Learning Algorithm | Supervised Machine Learning Algorithm |
| Algorithm Chosen: Random Forest Classifier |                                       |                                       |

According to the description and analysis performed above, we come to the following conclusion:

- 1) SVM performs well on small amount of data but has low accuracy as compared to random forest
- 2) Neural Networks gives optimized results but are time consuming and require large amount of data
- 3) Random Forest Algorithm has given the highest accuracy (0.9782430213464697), as compared to other two algorithms and are most accessible for tabular data.

Thus, we decided to go for random forest algorithm.

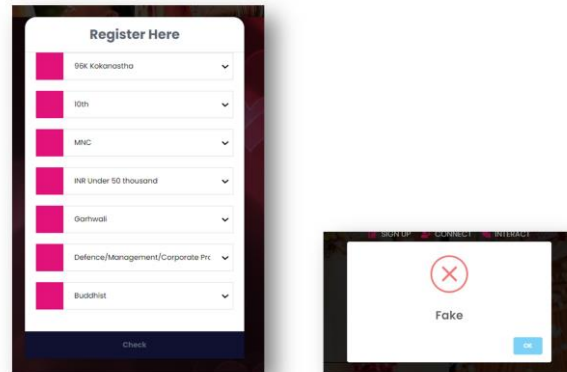
## VI. RESULTS

After the user fills in the required details, depending on the inputs provided, the profile is classified as Fake or Real

In (Fig. 17) of registration form, the user has entered the details which are all contradicting, thus it is declared as a fake profile:

- 1) Caste: 96K Kokanastha
- 2) Degree: 10th
- 3) Employment Sector: MNC
- 4) Income: INR Under 50 thousand
- 5) Mother Tongue: Garhwali

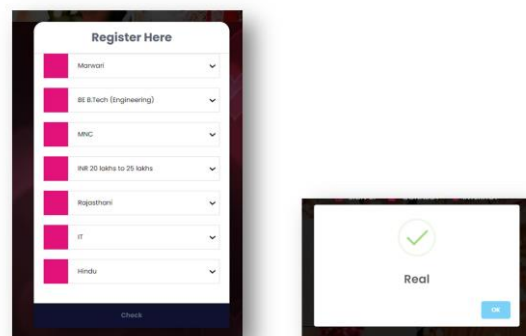
- 6) Occupation: Defence/Management/Corporate Professionals
- 7) Religion: Buddhist



**Fig. 16:** Details corresponding to Fake Profile

In (Fig. 18) of the registration form, the user has entered the details which are all authentic, thus it is declared as a real profile:

- 1) Caste: Marwari
- 2) Degree: BE B. Tech (Engineering)
- 3) Employment Sector: MNC
- 4) Income: INR 20 lakhs to 25 lakhs
- 5) Mother Tongue: Rajasthani
- 6) Occupation: IT
- 7) Religion: Hindu



**Fig. 17:** Details corresponding to Real Profile

## VII. CONCLUSION

Depending upon the details provided, our model is able to identify whether the profile is fake or real. The values which are taken into consideration influence the authenticity of the registered profile which will further help to reduce the number of fake profiles on matrimonial sites.

Though the project initially focusses on fake profile detection, its scope can be extended to various other features. According to the survey, most of the matrimonial sites are not user friendly and makes it difficult for the user to identify its

features and use them. To overcome this problem, UI/UX of existing websites can be made more interactive and user friendly. The recommendations provided by the existing systems keep flooding the inbox of the user. It is found that most of the profiles don't match the criteria provided by the user. This can be taken care by improving the quality of recommendation system of existing websites. Other security facilities such as Optical character Recognition (OCR), for validating the information provided by the user can also be incorporated which are not so common in existing systems.

## REFERENCES

- [1] Sneha Pingale, Chandrashekhar Kaushik. (2020). A Study on Digitalization of Matchmaking for Marriages in India.
- [2] Neethu Krishnan, Hemamala k, HariKrishnan.A.V. (2018). Study on the factors affecting the preferences of online matrimonial portals over traditional methods.
- [3] Shalinda Adikari, Kaushik Dutta. (2014). Identifying fake profiles in LinkedIn.
- [4] Vishal Sharma, Bonnie Nardi, Juliet Norton, A. M. Tsaasan. (2019). Towards Safe Spaces Online: A Study of Indian Matrimonial Websites.
- [5] Vaibhav Garg. (2019). Detecting Fake Profiles on Online Matrimony.
- [6] P. K. Roy, S. Chahar. (2020). Fake Profile Detection on Social Networking Websites: A Comprehensive Review.
- [7] Yasyn Elyusufi, Elyusufi Zakaria, M'hamed Ait Kbir. (2020). Social Networks Fake Profiles Detection Using Machine Learning Algorithms.
- [8] Ajesh F, Su Aswathy, Philip Felix, Jeyakrishnan V. (2021). A Hybrid Method for Fake Profile Detection in Social Network Using Artificial Intelligence.
- [9] Diksha Gupta. (2021). Detection of fake profiles using machine learning.
- [10] Sneha B Perla, Tejal P Sarvade, Balram C Chaurasiya. (2021). Detection of Fake Profiles on Online Matrimony Using Machine Learning and Deep Learning.
- [11] Liu, Chengwei, Yi-Chao Chan, Syed Hasnain Alam, Hao Fu. (2015). Financial Fraud Detection Model: Based on Random Forest." Econometrics: Econometric Model Construction.
- [12] Tamboli, Manisha & Ashish, G & Sahu, Madhaw & Sahu, Aish & Mohiddin, & Shukla, Arpita & Mohiddin, Dr-Md. (2021). Commercial Utilization of Chat-Bot Using Python.
- [13] <https://www.simplymarry.com/>