National College of Ireland

Dating Application Fraud Profile Detection and Analysis using Data Mining

MSc Research Project

MSc Research Project Data Analytics

Jinia Bhattacharya

Student ID: X23118890

School of Computing

National College of Ireland

Supervisor: Hamilton Niculescu

National College of Ireland

MSc Project SubmissionSheet

School of Computing

Student Name:	Jinia Bhattacharya		
StudentID:	x23118890		
Programme:	MSc In Data Analytics	**Year:**	2025
Module:	Research Project		
Lecturer:	Hamilton Niculescu		
SubmissionDue Date:	20th April 2025		
ProjectTitle:	Dating Application Fraud Profile Detection and Analysis using Data Mining		
WordCount:	7104	**Page Count:**	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Jinia Bhattacharya
Date:	20th April 2025

PLEASEREADTHEFOLLOWINGINSTRUCTIONSANDCHECKLIST

Attachacompletedcopyofthissheettoeachproject(includingmultiple copies)	□
AttachaMoodlesubmissionreceiptoftheonlineproject submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, bothforyourownreferenceandincaseaprojectislostormislaid.Itis notsufficienttokeepacopyoncomputer.	□

AssignmentsthataresubmittedtotheProgrammeCoordinatorOfficemustbeplaced into the assignment box located outside the office.

OfficeUseOnly	
Signature:	
Date:	
PenaltyApplied(ifapplicable):	

Dating Application Fraud Profile Detection and Analysis using Data Mining

Jinia Bhattacharya

X23118890

Abstract

With the popularity that online dating application has gained, safety of users has been compromised by the increase of fraudulent activities such as catfishing, identity theft, and financial scams. This work focuses on detecting and analysing fraudulent profiles using data mining and machine learning techniques. The K-Means clustering method of unsupervised learning techniques is used in this study to detect anomalies in user profiles. Text-mining approaches like sentiment analysis and topic modelling are employed to identify the deception patterns in the detected anomalies within the profile. In addition, Support Vector Machine (SVM) as classification models are used to forecast fraudulent profiles. Experimental results show that combining clustering, sentiment analysis, and classification is more accurate at detecting fraud resulting in higher precision in SVM. Future aspects will aim to improve the class-balance and integrate advanced NLP models including real-time datasets can make fraud detection in dating applications reliable.

Keywords: SVM, Random Forest classifier, DBSCAN,ODF, KMeans, Textblob, sentiment analysis, NLP, Topic Modelling, Data Mining, Text Mining

1 Introduction

Popularity of online dating is seen to be increasing in the recent past. In countries like USA, a good proportion of people have admitted to using such application Anderson et al. (2023).
The rising popularity also comes with some downsides where many users are left with negative experiences and many become victims of different scams as claimed by Anderson et al. (2023).
Different kind of scams sometimes involves catfishing, identity theft, bullying or financial fraud as mentioned by Lauckner et al. (2019). According to Bharne & Bhaladhare (2022),

most scams boils down to financial exploitation at some point and Bharne & Bhaladhare (2022) also discussed the societal implication of these kind of scams on the victims, where the victim might end up having major psychological and financial impact.
Keeping the societal implications in mind, a full scale research is needed in place to understand and detect the fraudulent profiles in these dating applications before hand by both the application end users and also their developers.
It is important how the fraudulent profile works which is why an investigation of similar fraudulent cases is done. Several similar frauds exist in different industries in a similar way like banking sector or social media. There are some studies like Shree et al. (2021) and Boulieris et al. (2024) that are done to mitigate these frauds on social media and banking. Another relevant study which is done to detect dating fraud is by Khan & Kamal (2023).
All these studies used Machine Learning Algorithms like classification and Natural language processing to detect and understand fraud and have been successful in that.
Shree et al. (2021) and Boulieris et al. (2024) employed hybrid models combining Natural Language Processing and Machine Learning approaches like classifiers i.e support vector machine to detect Fraud. Boulieris et al. (2024) also employed the usage of different Machine Learning algorithm like unsupervised learning along with the same. Thus it is very evident how Machine learning and NLP has helped with Fraud detection in different industries. **However, there are not enough studies which are in place for dating application fraud detection.**
This study aims to build upon the research done by Khan and Kamal (2023) by employing a hybrid approach of combining unsupervised learning and NLP on top of the supervised learning approach used in Khan and Kamal (2023) as the hybrid model has proven to be effectively detecting fraud in studies like Shree et al. (2021) and Boulieris et al. (2024) in other industries.
A public dataset of one of the dating application is picked for this research to analyse and detect fraud. The dataset belongs to okCupid dating application which contains textual and numerical information which is not labelled. This aligns with our idea to use hybrid model with NLP to understand the linguistic tone and sentiment of the possible fraudulent and non fraudulent profiles.
A predictive supervised learning algorithm like classifier is very important for Machine Learning algorithm and can be seen to be used in most studies like Khan and Kamal (2023) and Shree et al. (2021) , related to predicting fraud. Shree et al. (2021) employed the use of SVM which works well with high volume data. SVM seemingly an ideal choice for high volume data, could prove to be a good choice for predicting fraud.

The textual and numerical information present in the okCupid dataset is used to detect fraud by analysing unusual behaviour and forming outlier cluster and labelling them with unsupervised learning. **Unsupervised learning was important as the dataset needed to be labelled with some potential fraudulent profiles in order for the classifier to work** as used by Boulieris et al. (2024).
Therefore, KMeans is chosen for clustering and labelling the dataset by forming clusters as it is efficient with cluster formation based on outlier behaviour as claimed by studies like Oti et al. (2021).**Clustering detects and form an initial cluster of fraudulent data for the NLP to analyse fraudulent behaviour and for the classifier to refine the detection further.**

The hybrid models in studies like Boulieris et al. (2024) has also provided the motivation to employ the usage of NLP techniques like Sentiment Analysis on top of other machine learning models which can potentially boost the understanding of the detected fraudulent profiles.

While going through sentiment analysis approaches from studies like Kannappan (2023), Text Blob libaries for sentiment analysis is a predominant approach and advanced NLP tool like Topic Modelling provides detailed analysis and context on text based datasets as claimed by studies like Kherwa & Bansal (2018) is used. Both the NLP approaches help understand more about the fraudulent behaviour.
While Sentiment Analysis can provide the sentiment of the fraudulent and non fraudulent profiles, Topic Modelling can further break down the text to understand the themes. These NLP approaches can provide a detailed analysis on fraudulent linguistic patterns.

Taking into consideration the previous studies and approaches, this study aims to address the dating scams mentioned in Lauckner et al. (2019) and Bharne & Bhaladhare (2022) in two steps:

1. Detection of fraud from the Dataset
2. Analyze the linguistic Pattern of fraudulent profiles

The detection of fraud could be achieved by Machine learning algorithms (supervised and unsupervised learning) and the analysis of the frauds could be achieved using NLP (Sentiment Analysis and Topic modelling).

This study also aims to answer the below questions:

- How does sentiment analysis differentiate between the validation of fake and real dating profiles?
- What is the effect of clustering techniques on the detection of fake profiles?

2 Related Work

The rise of dating applications in countries like USA is rapidly increasing. One such study that quantified the amount of dating application usage in USA is Anderson et al. (2023).It maintains that three in every ten individual in U.S use dating applications and one in ten end up in a committed relationship with their dating matches from these applications. However, this study also reflects the overwhelming number of people (52%) that are left with negative experiences in these applications also highlights the requirement of safety. Researchers get an idea on the target vulnerable audience (women, elderly people of certain age) who can be subjected to these negative dating application experiences Anderson et al. (2023).
The number of negative responses has opened doors for more research on this area and provides enough reasons to investigate potential ways to prevent scams and work ways to make an impact on application experience positively.
The negative experiences highlighted in Anderson et al. (2023) is reemphasized by another study by Lauckner et al. (2019) which explores the types of scams and reemphasizes the figures provided by Anderson et al. (2023), and produces the fact that deception is indeed very common in the dating world and is commonly termed as ‘catfishing’ which has been to seen to take the form of financial exploitation.
Catfishing is done for several reasons and one of the main reasons is seen to be financial which is explained by Bharne & Bhaladhare (2022). According to Bharne and Bhaladhare (2022), there are different stages of online dating fraud. The first stage is profile creation

where scammers create fake profiles with attractive photos and false personal information to attract potential victims. The second phase is establishing trust by consistently projecting a reliable and compelling narrative, scammers create rapport and build a sense of trust and emotional attachment with the victim. The third phase is emotional engineering by using the trust therefore acquired; the frauds manipulate their goal emotions to feel love as well as dependence on their goal. The fourth phase is grooming. The grooming stage is one where the relationship deepens, and elements of sexual exploitation or coercion are introduced, drawing the victim in emotionally. The last and fifth stage is financial exploitation and it provides false information about crises or needs that can lead victims to provide financial support under pretences. **It is essential to understand these phases to design effective detection mechanisms.** Understanding the patterns and tactics of each stage will allow the study to formulate targeted strategies for early detection and prevention of fraudulent activity, protecting potential victims from emotional and financial impact.

Kristy et al. (2023) also states the tendency of the fraudulent behaviour of an over-exaggerated emotions to convince their victims of their virtual identity. It can be faking certain details or their appearance in their pictures or using luring words in their bio or texts to emotionally manipulate the victims.

Early detection is really important to prevent the frauds from happening. The operation of Dating application is similar to Social Media as both facilitates the profile browsing and profile interaction features. **There are previous studies on detecting fake or suspicious profiles in Social Media.** Shree et al. (2021) aims at identifying fake profiles of Instagram using Machine Learning. Machine Learning algorithm like support vector machine is combined with Natural Language Processing to find these fake profiles. From the study of Shree et al. (2021), it is established that the hybrid model of Machine learning and NLP is an effective approach for Fraud detection.

Another similar study by Khan and Kamal (2023) performed a very similar study with the same objective of finding fake profiles in dating applications. It has also employed the usage of machine learning classifiers to identify fake profiles. Unlike Shree et al. (2021), Khan and Kamal (2023) did not explore the possibility of hybrid model and usage of NLP to identify these fake profiles which may lead to missing subtle linguistic patterns that signal potentially fraudulent behaviour.

In contrast, Boulieris et al. (2024) combine the NLP with classification models subsequently achieving improved fraud detection in banking through the analysis of the textual data from financial transactions. They also highlight the significance of unsupervised learning for anomaly detection, which is applicable to datasets without labelled instances of fraud. While Boulieris et al. (2024) focus of their work is on the financial sector, the methods they propose (linking Natural Language Processing to supervised machine learning classifiers, along with unsupervised learning for detecting anomalies) are helpful for insights related to fake profiles detection in online dating realm. Thus, the combination of NLP techniques, along with machine learning classifiers, could supplement unsupervised learning methods to improve the accuracy and robustness of fake profile detection systems across domains.

As Boulieris et al. (2024) used unsupervised learning for fraud detection, it is important to understand the techniques of unsupervised learning that could be employed. Oti et al. (2021) provide an in-depth overview of one of the highly effective unsupervised learning approaches like K-means clustering algorithms and discusses its pros and cons.

Since Boulieris et al. (2024), Khan and Kamal (2023) and Shree et al. (2021), all of these employed classification models for fraud detection but these studies haven't discussed the challenges so it is important to address them.

The methodologies proposed in some papers like Boulieris et al. (2024) may be readily applicable to relatively structured datasets, but how universally applicable they are to the

current research based on unlabelled data, which does not have categories of domains, is still an open question. This scenario needs more investigation on detecting fake profiles using such clustering techniques without any labelled data.

For a further exploration of NLP and its applications, the study by Kannappan (2023) is referred. This research describes applying NLP techniques like TexBlob and machine learning to process and analyze the sentiment in the reviews the customers provide. This is useful and the same technique has also been applied to find out the Analytical sentiments on the texts in the bios which has been provided in the profiles of the fake and genuine profiles. Also, this study states the general architecture of sentiment analysis and highlights feature extraction and vectorization of text. The data is then numerated with text and the sentences using a count vectorizer for feature extraction Kannappan (2023). One of the most common methods of turning text into a vector is the Bag-of-Words.

Data Preprocessing of the textual data is also a very big part of using Machine learning prior to performing sentiment Analysis used in Kannappan (2023) .Different Data preprocessing concepts of textual data like tokenization and lemmatization are covered by Hazarika et al (2020).

From the above studies it is clearly established that most of the previous work employed NLP in the models they used to understand the profiles better and detect fraud.

Another method that is employed along with sentiment Analysis to have deeper understanding of the text and to identify underlying pattern is Topic Modelling as mentioned by Kherwa & Bansal (2018).

Overall, the above studies highlighted using classification models like SVM and clustering techniques like Kmeans along with NLP techniques like Sentiment Analysis, Topic modelling for an efficient fraud detection. The above studies cover fraud detection using hybrid model methodologies on different domains like social media and banking which works similarly like dating application. However, very little research has been done on fraud detection of dating application which opens door to a lot of researches on this area. This underexplored topic needs more focus and investigation just as efforts are made to detect and prevent fraud in banking and social media. With the rising popularity of dating application in the past few decades, addressing this negative impact has become significantly important. **Focusing on this under researched area of dating fraud detection opens potential option for greater innovation and overall significant contribution in the world of fraud detection.**

2.1 Summary of Findings

Firstly, the popularity of dating applications is discussed and then researchers analyze various dating frauds that came along, as well as the motives behind these frauds and some patterns of the possible fraudulent profiles. These helped people understand the area people needed to work on i.e. profile pictures, bios, etc. To work with profile information, researchers need a machine-learning approach to text analysis. Using sentiment analysis and NLP analysis appears to be a good option, considering the accuracy and efficiency demonstrated in the above papers, and researchers also recognize the need and application of unsupervised learning for the unlabelled dataset. Although many of these papers used similar procedures for different domains like social media or banking, very little research

has been conducted on the detection and analysis of Online Dating Fraud. Lastly, the importance of predictive models like classification is highlighted.

The combination of classifiers like SVM and NLP is motivated by papers like Shree et al. (2021) and Boulieris et al. (2024). The idea of implementing topic modelling for better understanding and dimensionality reduction is provided in the paper Kherwa & Bansal (2018) and Kmeans for clustering anomaly detection is motivated by Oti et al. (2021).

2.2 Justification of Research

The extensive use of online dating apps has definitely transformed the way meet people socially and with that comes the risk of getting scammed and the increase of fake profiles. These fake profiles dupes the identity of a real users to cause monetary and emotional damage as claimed by Bharne & Bhaladhare (2022). Anderson et al. (2023) paper helps in understanding the prevalence of online dating in the U.S., and demonstrates the need to study the fraud by quantifying the risks and negative experience associated with their use. For example, in their extensive research article, Bharne and Bhaladhare (2022), present a detailed overview of online dating fraud covering the methods used by such scams and the implications on their victims. Considering the scale and complexity of this problem, it is crucial to have reliable detection systems in place to protect users and keep dating zones safe. There are different methods already existing to identify fraudulent profiles, as mentioned above. Khan and Kamal (2023), applied machine learning classifiers and reached admirable accuracy in identifying fake profiles. Although mainly focused on studying structured and labelled data but it could not capture subtle textual signals in user profiles. In contrast, Boulieris et al. (2024) established that F2NLP is a hybrid of NLP with ML algorithms and to analyse linguistic patterns for fraud detection of banking services. It is common to see scammers using language-based strategies especially in dating applications to deceive their victims as Language is the only form of communication for a good amount of time between the end users and scammers. Despite these advancements, challenges still persist and more advanced research is needed for dating applications as there are very few research currently on this domain particularly with the detection of sophisticated scams that evolve over time. Hence, a multi-faceted solution involving machine learning, NLP, clubbed with anomaly detection should be used minimize the impact of online dating fraud.

3 Research Methodology

User information in dating platforms is significant in providing information that identifies an individual. But faux profiles walk a fine line with their overly emotive responses, unrealistic financial descriptions, and mixed-up language. To detect such cases, a machine learning-based workflow employing natural language processing, text mining, and clustering algorithms needs to be devised. Like Boulieris et al. (2024) has shown promising results with hybrid model in fraud detection, this study also employs a hybrid model with unsupervised learning for anomaly detection combined with NLP for sentiment analysis of the fraudulent profiles and classification algorithm for advanced fraud detection.

In order to employ the hybrid model, the data is collected from from an open source platform named Kaggle (https://www.kaggle.com/datasets/subhamyadav580/dating-site) and the dataset belongs to a popular Dating application named from OkCupid is used in this study. The Dataset has various information containing age, profession, education and height and there are essay fields where the users express themselves in their dating profiles. The dataset contains both Numerical and text fields where the numerical field mostly are fields like height, age and income and Fields like Job, Location and education are encoded and converted to numerical data for processing as these are crucial fields for deciding fraudulent profiles. Along with this, there are essays of text, that each user has written to describe themselves which this study makes use to analyse the intent and find deviation and abnormal behaviour in the text which fits into the previously found fraudulent behaviour or discover new pattern of fraudulent behaviour. Overall, this study focuses on finding the below objectives:

1. Detecting fraud (clustering and classification)
2. Understanding the underlying patterns of the detected fraudulent profiles.

The dataset used here is not a labelled dataset which means there are no labels for any suspicious profiles. The presence of unlabelled dataset requires unsupervised learning as mentioned in Boulieris et al. (2024) and the same is used in this study as the first step in the pipeline as mentioned in the flow diagram Figure 1.

The flow diagram of the approach is as attached below:

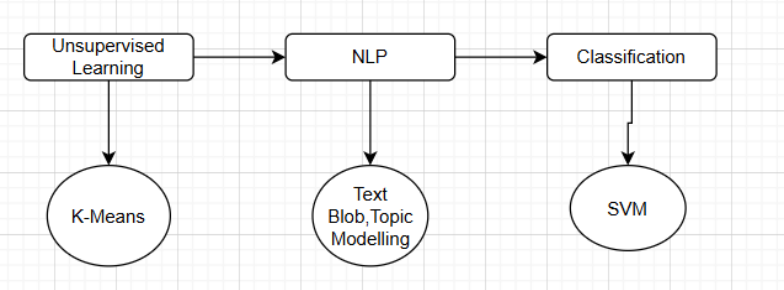

Figure 1 Model Architecture Diagram

The diagram in Figure 1 shows the model architecture of different steps undertaken for fraud detection in dating application.

Step 1: Unsupervised Learning -This is an important step employed for unlabelled dataset and has been an effective solution for finding fraudulent profiles in Boulieris et al. (2024) .Clustering groups similar data points and finds out outlier points. The outlier points are the possible fraudulent profiles.

Step 2: Natural Language Processing- As seen in most of the previous studies like Boulieris et al. (2024), Shree et al. (2021), NLP has contributed a lot in understanding behavioural pattern of the fraudulent profiles. This gives a better idea on the linguistic pattern of the fraudulent profiles. NLP is performed on the possible fraudulent profiles from step 1 clustering.

Step 3: Classification- Even though clustering is the first step in detecting the fraudulent profiles, there needs to be more refined and accurate detection of the model as sometimes not all anomalies are frauds according to Boulieris et al. (2024). The classification model used in this step learns about the fraudulent and non-fraudulent data from the results of the clustering model and detect the fraud based on those results.

The above steps undertaken helps address the issues highlighted by Lauckner et al., (2019) and Bharne & Bhaladhare (2022).For each of the step employed above, different technologies are used as shown in the Figure 1.

Different technologies, like Support Vector Machines (SVM), NLP, topic models, and clustering techniques like K-Means, are the possible solutions for the task. Shree et al., (2021), found that these combinations have shown powerful performance in fraud detection too, especially when the hybrid method is used to give better prediction accuracy.

-Researchers need to apply unsupervised learning and K-Means for initial clustering since the data set is not labelled.

-To resolve the problem of fraud mining, researchers apply text mining techniques like Text Blob, Topic Modelling as highlighted by Kherwa & Bansal (2018) and Hazarika et al. (2020)
- SVM for classification as shown in Figure 1.

Before the application of machine learning and text mining, the data needs to be processed as a part of the pipeline. Feature extraction is also required post Data pre-processing as a necessary step for performing the Machine Learning Algorithms as it converts the text to numerical value. The NLP requires feature extraction step to convert text to numerical data so that it can perform sentiment analysis and topic modelling and understand linguistic fraudulent pattern.

Using a structured pipeline a model is created to perform a correct classification:

Data Pre processing*:* Removal of stop words and punctuation, tokenization, lemmatization, and missing value handling and removing duplicates

Feature Extraction: Count Vectorizer is performed for the Model, where feature extraction is carried out.

Clustering: Model uses K-Means for clustering

Text Mining & Sentiment Analysis*:* Fraud indicators are analyzed through TextBlob.

Classification*:* SVM profiles fake and real classification is done.

With this methodology applied, numbers and text is thoroughly examined, leading to an efficient date application fake-profile detection method.

In the dataset used, there are numerical and textual data present which are combined for performing Machine Learning algorithms to it. From papers like Bharne & Bhaladhare (2022), some traits in fake profiles were identified i.e most of them would try to convince their victims and sell an idea of having an above average Personal Information like Salary or education or exaggerated emotions.

This study helps detect those fraudulent profiles, cross validate and provide more details on the claims by Bharne & Bhaladhare (2022) on the fraudulent profile traits and discusses those patterns in detail using the Pipeline mentioned above. The above discussed pipeline is visually explained in more detail in Research Architecture Diagram in Figure 2.

3.1 Research Architecture

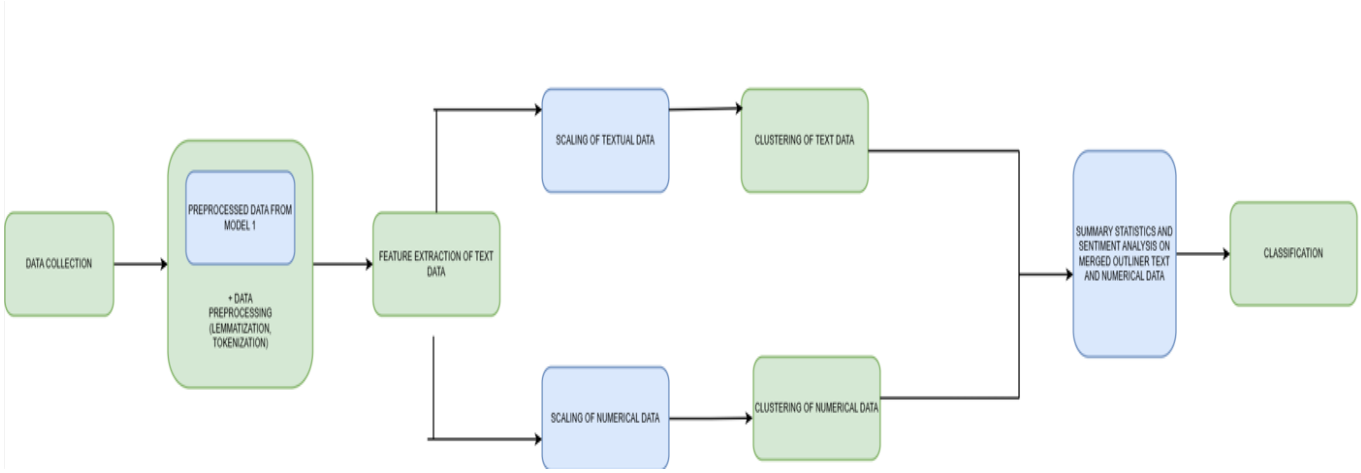

Figure 2: Model Flow diagram

The Architecture diagram shows the Data Collection as the first step followed by Data pre-processing, Feature extraction, Scaling, Clustering, Sentiment analysis and Classification. Model performs advanced data preparation like lemmatization and tokenization as motivated by Hazarika et al. (2020) as a data pre processing step for better normalization of the text. The numerical data is also pre-processed along with the textual data. The pipeline begins with data collection and pre-processing, where they are fine-tuned and is further augmented. Here, usage of textual data wherein feature extraction is conducted to produce vector representations and prepare it for machine learning algorithm, and both textual and numerical data are scaled.

The Model performs clustering independently for textual data and numerical data, treating them as different components. The clustering is done on two different types of data which are present in the dataset i.e Numerical and Textual Dataset. This is performed in order to gain understand the fraudulent profiles detected and characteristics derived from each of the dataset and the contribution of each of these clusters in the Algorithm. After the clustering, summary statistics and sentiment analysis applied to the outlier data helps understand what suspicious accounts exhibit.

The classification is then done using ML approaches such as SVM to classify the profiles as being fraudulent or legitimate. The modular structure of the Model achieves a higher degree of fraud detection accuracy as compared to other models, due to its further data pre-processing and separate clustering methods. Advanced procedures are used to process and classify text, this method helps to enhance fraud detection efficiency for dating applications.

3.2 Technologies Applied

Step	Algorithm	Libraries
Clustering	Kmeans	SKLearn,Numpy
NLP		TextBlob,SKLearn,NLTK
Classification	SVM	SKLearn,Numpy

Fig 3: Cluster-Based Fake Profile Detection

As mentioned in Figure 3, K-Means Clustering assigns profiles to fixed clusters. Therefore, it is very effective in differentiating the patterns of fake and real profiles.

SVM Algorithm is used for classification as it has great interpretability and is quite efficient for classification. Some of the supporting libraries for each of the steps are: NumPy which is normally used for numerical computation. Numpy acts as the storage room for the data to be computed for both the Machine learning Algorithms like clustering and classification. Sklearn library is also used in both the algorithms as it is a powerful library and supports the Machine learning algorithms with less effort as it has already built implementations embedded in it which made the computation of the algorithms easier because of the prebuilt codes.

TextBlob library is being used for sentiment analysis as it is known to be efficient for processing textual data and NLTK (Natural Language Toolkit**)** is an old library which has been used for NLP processing and it comes with many advanced and a number of features like tokenization, stopword removal and lemmatization which makes it one of the most popular library to be used for NLP.

4 Design Specification

Step	**Model**
1.Data Collection	Collected Data from Kaggle
2.Text Pre-processing	Removal of unwanted words, spaces, and special characters
	Tokenization
	Lemmatization

2.Numerical Pre-processing	Removal of missing values and duplicates
3.Feature Extraction	Count Vectorizer for numerating text data
4.Clustering	Finding outliers in the dataset
5.NLP	Sentiment Analysis and Topic Modelling applied to the outliers and non-outliers of step4
6.Classification	For refined detection of Fraud based on the clustered outliers and NLP data output

Table 1: Model Designs

The Table 1 further explains the 6 steps mentioned in Figure 2.Each step is discussed in detail later in the implementation section.

5 Implementation

The six steps mentioned in Figure 2 flow diagram and Table 1 of the Model is explained in detail below:

5.1 Dataset Collection

The data is picked from an open source website named Kaggle (https://www.kaggle.com/datasets/subhamyadav580/dating-site) which belongs to an U.S based dating application named OkCupid. It contains profile information of more than 50,000 individuals and has 31 attributes. Some distinguished profiles information that are used to detect fake profiles are age, income, height, location, job and education and essays. Data fields like profile bio(Text essays),Job, education, height mentioned above can have significant contribution in identifying fraudulent profiles as Bharne & Bhaladhare (2022) that the fraudsters might try to exaggerate this information to convince victim of their reliability financially and emotionally. For example, a profile with a good financial and educational background and a profile with more emotional availability are often more reliable as compared to others.

5.2 Data Pre-Processing

In order to perform Machine learning Algorithm and NLP to detect and define fraud, Data pre-processing is a mandatory step as the dataset picked already have textual data and numerical data. The textual information need to be cleaned and processed and they are unstructured and inconsistent which makes it difficult for the algorithms to process. The numerical data also needs cleaning to make sure of encoding and cleaning of the values. This is a mandatory step to perform the machine Learning algorithms and NLP as also seen in studies before Shree et al. (2021).

Text Data: The multiple essay fields in the data are combined, stripped of the extra spaces and special characters that are not alphanumeric and duplicates are dropped from the dataset. Stop words are also removed from the combined essays.

Numeric Data: The numerical data is prepared by removing all the missing values or null characters.

Encoding of Numeric Fields: Fields like Job, education and location are **encoded** with values including missing values. Encoding them helps the algorithm learn more about data allowing it for better processing (i.e clustering). These features are applied to combat vast range of values and exclude biases in the algorithm.

Few extra steps were undertaken to pre-process the text data.

Text data: Stop words (‘and’,’the’), which do not significantly contribute to the meaning of the sentence, are first removed. Then, **Lemmatization** is used, which is a process for understanding the context of a word and converting it to the dictionary form of the word. According to Nikiforos *et al.* (2021), for Lemmatization, libraries like Open Multilingual WordNet (OMW) and WordNet are used, which are known for grouping synonyms and providing meanings to the words. Before Lemmatization, the process of tokenization is used to extracts few meaningful words from a sentence and those words are lemmatized and joined back into sentences to provide more meaningful input. For tokenization , library like Punkt, which is an unsupervised model of tokenization, is used Abraham (2024).

5.3 Data Analysis

Before data processing, attribute correlation needs to be visualized and analyzed before the clustering of the data and a heat map is used to understand the correlation between them. The correlation matrix helps understand the relationship between the data points, which helps with the clustering later.

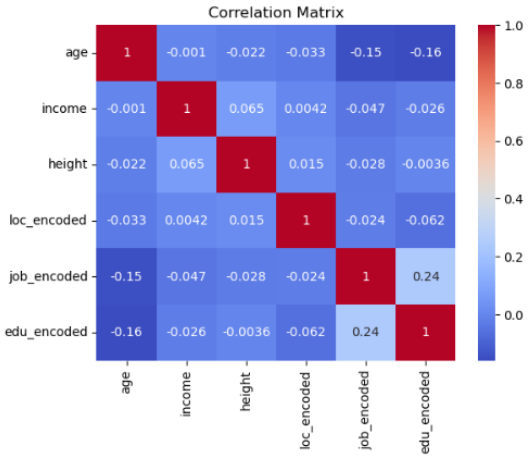

Figure 4: Heatmap of data

A correlation value in the range between -1 to 0 means the variables are negatively correlated and 0 to 1 mean those are positively correlated Investopedia (2023). The strongest correlation is between education and Job as seen in Figure 4, which is expected and rest of the values don't have any signification correlation amongst them. **Lower correlations can benefit clustering and ensures each feature contributions to the clustering process.**

5.4 Feature extraction

Feature extraction is an important step for Algorithm to process the high dimension textual data. It converts the text data data to numerical format which makes it easy for machine learning algorithms to process the data.

For the feature extraction, count vectorizer is used which forms a vector for the frequency of the words used in the document. This is helpful especially for unstructured data to be prepared into a machine learning ready data Motitswane (2023). It is a simple form of understanding the frequency of words in a document and the max features is being set 1000, thus making sure only 1000 more frequently used words is picked up from the documents. The Lemmatized text output from the Data preprocessing process is being fed into the count vectorizer. The result is an array (Dense Matrix) with rows as the number of data points and columns with the 1000 frequently most used words for each of the textual dataset.

After the feature extraction, both the numerical and textual variable are in numerical format that need to be standardized and scaled for the ease of machine learning algorithm so that the large variation in data is normalized. This is done to reduce bias in the algorithm with extreme values.

5.5 Clustering

After the data is cleaned, pre-processed , scaled and converted into numerical data, clustering approach like Kmeans is applied to the data to form clusters of the data and form a cluster for probable fraudulent and non-fraudulent profile data.

K Means is used to divide the dataset into 5 possible clusters Yildirim et al. (2021). It relies on the concept of Euclidean concept to form these 5 clusters. The minimum distance between the data point and cluster has been set by keeping it at 1. The 5 % of data points which are farthest from the clusters are considered as outliers. The outlier detection for both the numerical and textual data are carried out separately for each dataset and then the final outlier results are combined to get a total of 5812 outliers.

Clustering helps identify dissimilar behavior which could potentially be fraudulent profiles thus reducing the possible threats highlighted by Bharne and Bhaladhare (2022).

5.6 NLP

After clustering, NLP is applied to the clustered data. NLP approach like Text Blob is used for sentiment analysis. As mentioned before, this is helpful for understanding the overall sentiment of the fraudulent profiles and genuine profiles and the difference in the sentiments gives more details on their linguistic patterns.

This helps the end users distinguish fraudster behaviour from genuine profiles and in a way that helps with detecting suspicious behaviour from the user end.

The fraudulent outlier behaviour is further analysed and put into more context with the help of Topic Modelling which divides the text essays in the profiles into different topics or themes and put a meaning for each of the topics. **Putting context to the text by Topic Modelling helps identify the linguistic preference of the fraudulent and genuine profiles** and also helps with distinguishing the topics the profiles are more inclined towards. This gives more clarity in understanding and flagging the fraudulent account thereby upgrading the security of the platform.

5.7 Classification

This is the last step of the flow and is performed to enhance the fraud detection that are already identified by the clustering approach. The input data is the data on which clustering and NLP approach like Topic Modelling is performed. The output resultant data of

the clustering and topic modelling serves as an input for the model and the model is trained and tested on the same. The model learns about detecting fraudulent profiles from the clusters and the topics created in the topic modelling step help with dimensionality reduction of the model and ease of processing of the text data.

SVM is great in high dimensional data because, in using kernel functions (like RBF) they transform the data into higher dimensional space and linear reparability is easier to achieve Aiad *et al.* (2021). SVC, therefore, finds the optimal hyper planes for separating the data points, even in the case of non-linear scenarios, by finding the maximum gap between the classes.

6 Results and Evaluation

Clustering Analysis

As mentioned before, the Clustering with Kmeans on the dating application data yielded a combined 5812 outliers for both textual and numerical data set. Silhouette metric, which measures the efficiency of the clusters, is used to measure the textual and numerical outlier cluster individually. A sample dataset of 6000 is created for faster computation on each of the dataset.

The Silhouette score for numerical dataset and textual dataset are 0.20 and 0.0249 as seen in Figure 5.While 0.20 is not an ideal score, High dimension of the text data is a possible reason for the same. The silhouette score implies that the numerical cluster have higher influence on the performance of the classification as compared to the text due to weaker clustering.

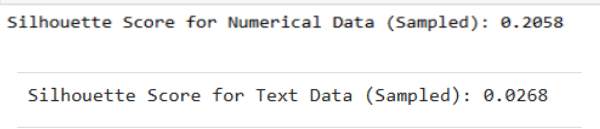

Fig 5 Metric for clusters

The outlier and non-outlier clustered data are analysed through box plot by comparing their summaries like mean, standard deviation and variability values.

According to Figure 6, most of the fields have higher variability in the outliers (possible fraudulent) compared to the non-outliers (genuine). **Higher Variability in the profile values indicates that sometimes fraudulent (outlier) has more extreme and inconsistent information i.e., extreme values on education, job, income** with their averages or minimum, maximum more or less than average genuine profiles in their profiles.

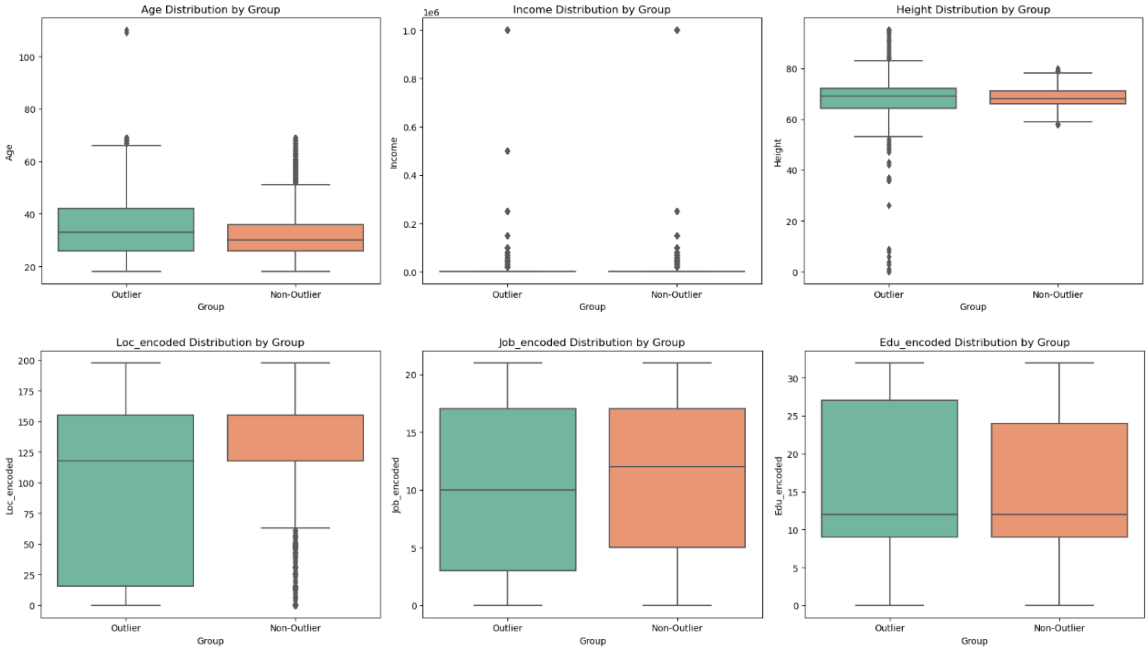

Fig 6 Boxplots for cluster summaries for Model

The conclusion derived from the cluster summaries definitely addresses to one of the goals of the study which is to understand the underlying fraudulent patterns from their profile information. This observation aligns with the claim by Bharne & Bhaladhare (2022) on the exaggerated and inconsistent profile information on fields like Job and Education. **This is further studied in detail in the next section with NLP**.

Evaluation:

The results provide insights on the difficulties of effectively clustering fake profiles.. KMeans doesn't yield the most efficient cluster of textual data but is computationally very efficient. KMeans can process large amount of data in less time although the efficiency of the cluster of certain data types is questionable. The results shown in Figure 5 show that clustering on numerical data is better than clustering on textual data, thus implying that fraud detection is more reliant on patterns in structured data than (textual) signals.

K-Means organizes five clusters, one that examines only the numerical values and a second one that inspects the texts individually. Although computationally efficient, it has a poor Silhouette score for text 0.0249 and decent score of 0.20 for numerical data as shown in Figure 5. In the end, clustering helps with foundational fraud detection, but is most effective as a complement to classification models.

NLP on Clustered Dataset

The sentiment analysis score for non-outliers is 0.17 as compared to the 0.16 in the outliers. Although the average difference isn't much high, it depicts slightly higher positive emotion in the outliers. The higher standard deviation amongst outliers to non-outliers indicates wider sentiment variation in the same.

Analyzing the sentiment features that are present or absent in real and fake profiles, researchers can validate authenticity using sentiment analysis. e. K-Means identifies outliers possess a wider variability in sentiment, indicating inconsistency. The SVM classification methods improve the detection through the help of sentiment & clustering-based analysis to achieve balanced fraud detection. Thus, sentiment-based content filtering boosts authenticity detection across dating platforms.

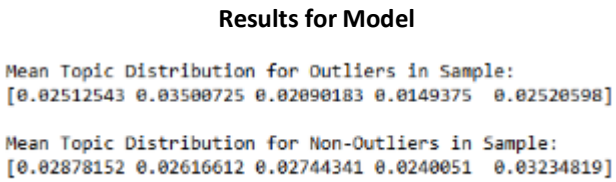

Fig 7: Topic Distribution

The distribution of each of the 5 topics (preference, relationship, hobby, personal adjective and social connections) is measured for both outlier and non-outlier datasets. The non-outliers exhibit higher average mean value for all other topics, indicating better performance and **more engagement with more diverse topics outside of relationships in genuine profiles**. Topics related to Family, friends and social connection (Topic 4) has the highest mean value 0.032 and 0.028 **, indicating that genuine profiles are more likely to choose or put topics on social connection** and preferences in their profile bio data.

Outliers with a higher mean value 0.035 for topic 2 compared to non-outliers, suggests more affiliation towards topics on relationship and enjoyment, showing more skew towards superficial topics .**Outliers are less likely to discuss on family, social life and preferences as compared to non-outliers**. Overall, most of the topic distribution values are higher in non-outliers which means genuine profile has more diverse context with more details, whereas outliers have less focused text. Thus, profiles with less topic distribution can be a potential suspicious profile.

Profiles that engage with authentic diverse topics are more reliable than the others.

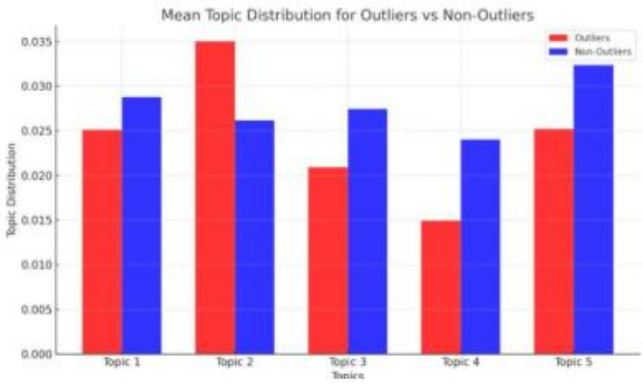

Fig 8: Topic Distribution graph

Evaluation:

This topic modelling shows difference in the engagement level for genuine and fraudulent profiles. Authentic profiles are seen to be taking more interest in multiple topics like Family, hobbies, etc while fraudulent profiles is more inclined towards a few selective shallow topics, such as relationships. Less variation in the topics many a times suggests a lack of genuine intent and inclines more towards deception. **Thus analyzing topic distribution is actually important in distinguishing between fake and genuine user profiles.**

Classification

Support Vector classifier is used as it can handle high dimension data. SVM is great in high dimensional data because, in using kernel functions (like RBF) they transform the data into higher dimensional space and linear separability is easier to achieve Aiad *et al.* (2021). SVM is a good choice for complex, high dimension data with fewer assumptions on data distribution and provides robustness of classifying data even when data is very imbalanced.

The input data account for 59948 non-outlier to 5810 outlier data. Both the text and numerical data are scaled and labelled with one hot encoding before training the model. The two different clusters i.e non-outlier and the actual outlier data detected from KMeans have been combined and split into Training and test data in a ratio of 70 to 30 as per standard rule and the model is trained with the training data and tested on the test data. The evaluation is based on Accuracy, Precision, Recall and F1 score.

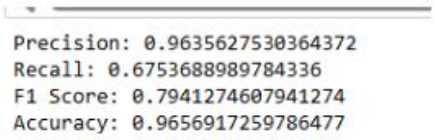

Fig 9 Classification result

Evaluation:

The high accuracy is because of lesser samples of outliers and is an expected outcome of the problem which could happen due to class imbalance as there was lesser number of samples for outliers.

Precision detects how many detected fraud cases are fraudulent, and decreases the amount of false positives Vorobyev *et al.* (2022).The Model has a good precision score which means it correctly identifies the fraud and has less false positive and it is important for the model as flagging a legitimate profiles could result in bad user experience and can create problem for the stakeholders. The recall 0.67 is an average portion of outlier (actual positive cases) that is detected which means some outliers could be missed. The model is thus trained with the results from the clustering data to detect fraudulent profiles efficiently and attempts to make predictions of fraudulent profiles.

There is a trade-off between precision and recall which is something commonly observed in the classification models. As SVM has a larger dataset, its precision is improved and recall is lower. A major issue is class imbalance, which influences the recall. Fusion of multiple models and better-balanced datasets can improve the outcome of fraud detection.

6.5 Discussion

The study provides evidence demonstrating the usefulness of combined machine learning and NLP techniques for detection of fraudulent profiles in dating applications. There are several **Findings that are observed in the study i.**e.,

-K-Means clustering methods were able to detect the 5812 anomalies.

-Patterns from fraudulent Clustered Anomalies were Inconsistent profile information and more extreme values in fields like Job, Education, etc

-Fake profiles were characterized by overly positive emotional expression as found in sentiment analysis

-topic modelling - fewer coverage of talking subjects of fake accounts with Lack of Investment in Diverse Topics like Family, hobbies, etc and Sticking to only limited topics like relationship

-Among the classification models, SVM performed with high precision and detected 1,165 fraud profiles successfully and 16,176 genuine profiles correctly as shown in Figure 10 of Confusion Matrix. Unfortunately, class imbalance and inefficiency in textual clustering was noted to negatively affect model accuracy. Combination of clustering, classification and NLP are needed to improve fraud detection, according to the research. **Deployment of advanced deep learning methods, optimized class balancing strategies, and real-time data validation techniques** should be explored in future work to improve accuracy of detection thereby providing safer online dating experiences for users.

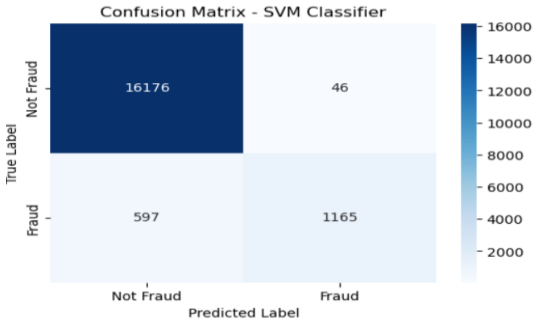

Fig 10 Confusion Matrix

6 Conclusion and Future Work

Implementation

This study successfully implements data-mining approaches using Python to identify and detect fraudulent profiles. It uses the following implementations:

Text vectorizer using Count Vectorization and clustering is performed using K-Means. Sentiment analysis is performed by using TextBlob that helps identify misleading emotions in fake profiles and for classification SVM Model is used. These implementations help answer the below research questions.

How does Natural language processing differentiate between fake and real datingprofiles?

The textual content of the dating profiles is put through a thorough analysis employing methods like sentiment analysis and cluster analysis for better understanding of the linguistic patterns that correlate with deception. The cluster analysis showed inconsistent profile information and the sentiment polarity scores obtained using TextBlob indicates

exaggerated positive emotions from the fake profiles which match with the deceitful type of behaviour. In addition, topic modelling shows that the fake profiles tend to focuses more on shallow topics related to relationships, while genuine profiles discusses on a variety of different topics. In this study, based on the above patterns, which were established by the clustering results and sentiment analysis, Accuracy of the SVM classifier is 96 in detecting fake profiles while the recall is affected due to the presence of class imbalance meaning the model performed slightly better at detecting genuine profiles than the fraudulent ones. The study touches on these limitations but rigorous resampling methods or more hybrid model strategies is not discussed in detail to alleviate them and the study doesn't discuss incorporating more sophisticated NLP methods like transformer-based models that could improve fraud detection.

What is the effect of clustering techniques on fake profile detection?
This study explores the usage of K-Means clustering in unsupervised fraud detection. Although it successfully creates decent fraudulent or outlier cluster with numerical data but it struggles to form meaningful clusters with text use-cases. Clustering helps detect the intial and usual suspects which don't align with regular behaviour. The presence of clusters helps the supervised learning model with the basic understanding of the fraudulent profiles which is largely helpful when the dataset is unlabelled. It also identifies the concealed patterns and is also largely helpful for evolving fraudulent techniques and on top of that clustering reduces the false positive in the classification model. While the study shows the impact of clustering on dating fraud detection setting, transformer based embeddings could be used for more contextual understanding and hyperparameter setting could be experimented with.

Contributions

This study introduces the integration of unsupervised clustering, sentiment analysis, and machine learning classification to develop a hybrid framework for fraud detection. Deceptive behaviours in dating profiles are successfully identified using NLP based text analysis. The identification of these patterns with the help of this hybrid model provides a new angle in detecting fraudulent activities which has not been explored before. This study puts the efficiencies of models like SVM and K-Means into test and examines their effectiveness for detecting frauds. The accuracy of fake profile identification is improved by leveraging the hybrid model and by integrating a layer of supervised learning and NLP. This dimension of work is a fairly new addition to dating fraud identification.

Real world use case

This study explores a new approach of combining unsupervised learning, NLP and supervised learning approach to detect and understand fraudulent profiles in a better way. This hybrid model experiment fairly resulted in identifying fraudulent behaviour and detecting them in a better way. The introduction of unsupervised learning helps the model identify fraud without needing labels for the dataset as compared to the traditional way of detecting fraud in dating websites as implemented in Khan, A. & Kamal, A. (2023) .Adding supervised learning refines improve precision in identifying fraud. Unlike previous models used in Khan, A. & Kamal, A. (2023) some hidden fraudulent patterns are also identified. The model is adaptive over time making it more dynamic than the traditional models. Overall, this study benefits both the end user and business in identifying fraud. The fraudulent patterns detected here would be beneficial for end users and fraud detection will help raise a red flag for business and application owners.

This study thus addresses the some of the negative experiences shown statistically by Anderson et al. (2023) to scams like catfishing as discussed by Lauckner et al. (2019).This study also aligns with the claims made by Bharne & Bhaladhare, (2022) and Kristy et al. (2023) on the fraudulent behaviour of emotionally manipulating the victims with false or inconsistent profile information and exaggerated emotions on profile essays.

This study also outperformed the supervised learning model to detect fraud in dating application by Khan and Kamal (2023) and showed that the introduction of hybrid model required and can serve multipurpose.

Limitations

A major limitation of this study is the class imbalance which is due to a smaller number of fraudulent cases found in the cluster. This scenario might have caused bias toward real profiles and affected the accuracy percentage. In addition, the study also suffers with the limitation on the textual clustering. Sentiment analysis does not provide fine-grained analysis of the linguistic pattern but lays out some generalised deception patterns. On resolving these issues, the model could perform better and prove to be more reliable for fraud detection.

Future work

- Methods such as SMOTE or (weighted loss functions) which is mainly used to balance class could work to increase recall.

- deep learning-based NLP models like BERT or GPT-based transformers could be used for Text fraud detection
- Live dating app datasets can improve the generalizability and the applicability of the models.

The current study although provides a solid base to identify fraudulent profiles in an online dating application but required more fine-tuning to achieve better ,balanced and less biased accuracy and could be used for real world application.

Ethical concern

One of the ethical concerns that rise from the above study is the dataset. The ownership of the dataset should lie with the users and their consent should be taken before any research. The dataset used here do not have any personal information like name or address of any user and it is taken from an open-source platform thus addressing the first ethical concern.

The major ethical concern that can rise is the results of the algorithm can be misleading especially if it falsely detects a genuine profile, so it is really important to have the false positive cases as minimum as possible. The False positive identified in this study is only 46 which is low which means very few legitimate profiles were wrongly tagged. However further research on this area can even further lower this count.

References

Abraham, A., 2024. Automatisierte Prüfung von Quellenangaben in studentischen Arbeiten (Doctoral dissertation, Universität Rostock).

Aiad, B.A.E., Zarif, K.B., Gadallah, Z.M. and Abd EL-kareem, H., 2021, August. '*Support vector machine kernel functions comparison'*. In The International Undergraduate Research Conference (Vol. 5, No. 5, pp. 84-88). The Military Technical College.

Anderson, M., Vogels, E.A. & Perrin, A., 2023. Key findings about online dating in the U.S. *Pew Research Center*. Available at: https://www.pewresearch.org/short-reads/2023/02/02/key- findings-about-online-dating-in-the-u-s/

Bharne, S. & Bhaladhare, P. (2022) 'Investigating Online Dating Fraud: An Extensive Review and Analysis', *2022 International Conference on Recent Trends in Microelectronics, Automation, Computing and Communications Systems (ICMACC)*, Hyderabad, India, pp. 141–147. DOI: 10.1109/ICMACC54824.2022.10093271.

Bhattacharjee, P. and Mitra, P., 2021. 'A survey of density based clustering algorithms'. Frontiers of Computer Science, 15, pp.1-27.

Boulieris, P. *et al.* (2024) 'Fraud detection with NLP', *Machine Learning*, 113, pp. 5087–5108. doi: 10.1007/s10994-023-06354-5.

Brindha, K. and Ramadevi, E., 2023. 'Twitter Sentiment Analysis For Feature Extraction Using Support Vector Machine (SVM) With TF-IDF'. Journal of Survey in Fisheries Sciences, pp.3575-3583.

Couch, D., Liamputtong, P. & Pitts, M. (2011) 'Online daters and the use of technology for surveillance and risk management', *International Journal of Emerging Technologies and Society*, 9(2), pp. 116–134.

Hazarika, D., Konwar, G., Deb, S. and Bora, D.J. (2020) 'Sentiment analysis on Twitter by using TextBlob for NLP ', *The International Conference on Research in Management & Technovation 2020*. doi: 10.15439/2020KM20

Nickolas, S. (2024) *What does it mean if the correlation coefficient is positive, negative, or zero?* Investopedia. Updated 26 December. Available at: https://www.investopedia.com/terms/c/correlationcoefficient.asp

Kannappan, S. (2023) 'Sentiment analysis using NLP and machine learning', *Journal of Data Acquisition and Processing*, 38(2), pp. 520-526. doi: 10.5281/zenodo.7766376

Khan, A. and Kamal, A. (2023) ‘Fake profile detection on dating websites using machine learning’, *International Journal of Advance Research and Innovation*, 11(2), pp. 21–25. doi: 10.51976/ijari.1122303.

Kherwa, P. and Bansal, P. (2018) 'Topic modeling: A comprehensive review', *ICST Transactions on Scalable Information Systems*, 7(24), pp. 159623. doi: 10.4108/eai.13-7-2018.159623

Kristy, A., Krisdinanto, N. & Akhsaniyah, A. (2023) 'Two Face Personality in Identity Falsification and Catfishing Behavior on Online Dating Tinder', *Communicatus Jurnal Ilmu Komunikasi*, 7(1), pp. 1–20. DOI: 10.15575/cjik.v7i1.26102.

Lauckner, C. et al. (2019) ‘“Catfishing,” cyberbullying, and coercion: An exploration of the risks associated with dating app use among rural sexual minority males’, *Journal of Gay & Lesbian Mental Health*. doi: 10.1080/19359705.2019.1587729.

Morstatter, F., Wu, L., Nazer, T.H., Carley, K.M. & Liu, H. (2016) 'A new approach to bot detection: Striking the balance between precision and recall', *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, San Francisco, CA, USA, pp. 533–540. DOI: 10.1109/ASONAM.2016.7752287.

Motitswane, O.G., 2023. Machine learning and deep learning techniques for natural language processing with application to audio recordings (Doctoral dissertation, North-West University (South Africa))

Moulavi, D. et al. (2014) ‘Density-based clustering validation’, *SIAM International Conference on Data Mining (SDM)*. doi: 10.1137/1.9781611973440.96.

Nesvijevskaia, A., Ouillade, S., Guilmin, P. and Zucker, J.D. (2021) ‘The accuracy versus interpretability trade-off in fraud detection model’, *Data & Policy*, 3, p. e12.

Nikiforos, M.N. et al. (2021) ‘The modern Greek language on the social web: A survey of data sets and mining applications’, *Data*, 6(5), p. 52

Oti, E.U., Olusola, M.O., Eze, F.C. and Enogwe, S.U. (2021) 'Comprehensive review of K-means clustering algorithms', *International Journal of Advances in Scientific Research and Engineering*, 7(8), pp. 64-69. doi: 10.31695/IJASRE.2021.34050

Prasetiyowati, M.I., Maulidevi, N.U. and Surendro, K. (2022) ‘The accuracy of Random Forest performance can be improved by conducting a feature selection with a balancing strategy’, *PeerJ Computer Science*, 8, p. e1041. doi: 10.7717/peerj-cs.1041

Robertson, S.E. (2004) ‘Understanding inverse document frequency: On theoretical arguments for IDF’, *Journal of Documentation*, 60(5), pp. 503–520. doi: 10.1108/00220410410560582

Rodríguez, J.F., Papale, M., Carminati, M. and Zanero, S. (2022) ‘A NLP approach for financial fraud detection’, *ITASEC’22: Italian Conference on Cybersecurity*, 20–23 June, Rome, Italy

Shree, S.S., Subhiksha, C. & Subhashini, R. (2021) 'Prediction of fake Instagram profiles using machine learning'. *SSRN*. Available at: https://ssrn.com/abstract=3802584

Sowmya, P. and Chatterjee, M. (2020) 'Detection of fake and clone accounts in Twitter using classification and distance measure algorithms', *2020 International Conference on Communication and Signal Processing (ICCSP)*, Chennai, India, pp. 0067-0070. IEEE. doi: 10.1109/ICCSP48568.2020.9182353

Talbia, D., Daouib, Z. and Gala, Z. (2024) 'Unsupervised machine learning-based clustering of high- frequency radio channel properties: analysis of sector communication efficiency', *Procedia Computer Science*, 238, pp. 306–313

Vorobyev, I. and Krivitskaya, **A.** (2022) ‘Reducing false positives in bank anti-fraud systems based on rule induction in distributed tree-based models’, *Computers & Security*, 120, p. 102786. doi: 10.1016/j.cose.2022.102786

Yazici, İ., Shayea, I. and Din, J. (2023) 'A survey of applications of artificial intelligence and machine learning in future mobile networks-enabled systems', *Engineering Science and Technology, an International Journal*, 44, p. 101455. doi: 10.1016/j.jestch.2023.101455

Yildirim, M.E., Kaya, M. and FurkanInce, I. (2022) ‘A case study: Unsupervised approach for tourist profile analysis by k-means clustering in Turkey’, *Journal of Internet Computing and Services*, 23(1), pp. 11–17