Detection of Deepfakes

Shivin Agarwal, 231110048, shivin23@iitk.ac.in
Lt cdr Sunil Kumar, 231110025, sunilklc23@iitk.ac.in
Yuvraj Raghuvanshi, 241110084, yuvrajpr24@iitk.ac.in
Group Number: 12

CS685: Data Mining

Abstract

Deepfake technology has raised significant ethical, social, and security concerns in the digital era. Our project aims to develop a web-based platform where users can register, access videos, and review them to determine whether they are authentic or deepfakes. By leveraging user feedback, the system collects valuable insights into patterns and behaviors associated with deepfake detection. The aggregated reviews are analyzed to identify trends, accuracy rates, and biases in user judgments, providing actionable insights into the efficacy of human evaluation in deepfake detection. This analysis serves as a foundation for improving detection methodologies and enhancing public awareness of deepfake technologies.

Motivation

The advent of deepfake technology has marked a significant leap in the capabilities of artificial intelligence, especially in the realm of digital media. By using deep learning algorithms, deepfake technology allowed filmmakers and content creators to generate realistic alterations to videos, such as dubbing movies in different languages and adjusting lip sync to match the dubbed dialogue. This technology also provided new possibilities in the realm of virtual reality, gaming, and digital media production, helping to create more immersive experiences and improve accessibility. The ability to replicate the faces of actors for dubbing or animation could even reduce production costs, making content creation more efficient and versatile.

However, as deepfake technology has evolved, its misuse has raised serious ethical and societal concerns. Although originally designed with creative and entertainment goals in mind, deepfakes have been increasingly exploited for malicious purposes. One of the most catastrophic uses of deepfake technology is the creation of harmful and defamatory content, such as revenge porn, where an individual's likeness is manipulated to create explicit or degrading material without their consent. Such videos have devastating consequences, particularly for women, leading to harassment, emotional distress, and the destruction of personal reputations. For example, a report from MIT Technology Review (2021)[4] highlights the rising concern over deepfake revenge porn, shedding light on its destructive impact on victims' lives. Similarly, deepfake videos have been used to tarnish the reputations of public figures, such as celebrities and politicians[5], and manipulate public opinion.

Moreover, deepfakes have also emerged as a tool for financial manipulation, with the potential to affect

stock markets and influence corporate behaviour. A notorious example occurred with a bank manager in Hong Kong who lost $25 million because of a highly convincing deepfake phone call [3]. Deepfake videos can also be weaponized in political arenas to sway elections or fuel misinformation. One notable case involved a deepfake video of Ukrainian President Zelenskyy, which falsely portrayed him urging Ukrainians to lay down their arms, an attempt to create confusion and distrust during a time of geopolitical crisis. Such instances expose the alarming potential of deepfakes to undermine democracy, destabilize economies, and erode trust in institutions.

Additionally, deepfakes have been employed for personal gain in commercial advertising, with instances where the likenesses of prominent figures such as business tycoons and political leaders were used in fake endorsements. For example, deepfake videos featuring figures like Nita Ambani and Gautam Adani [2] were circulated in ads promoting dubious mobile gaming apps, leading to concerns about the ethical implications of using individuals' faces and voices for unauthorized marketing.

Despite the considerable risks associated with deepfakes, the technology itself is not inherently negative. As with any technological advancement, the key lies in how it is used. The ability to create hyper-realistic digital content has vast potential for positive applications, from enhancing creative works to providing new forms of communication. However, as deepfake technology continues to evolve, it becomes increasingly important for individuals to understand its implications and learn how to detect and combat malicious deepfakes. With the rapid rise of deepfake-related threats, public awareness and the development of detection tools are essential to mitigating the harmful effects of this technology. In this context, the development of effective deepfake detection methods is crucial not only for protecting individuals' reputations but also for safeguarding the integrity of digital media, political discourse, and financial markets.

1 Data Used

1.1 Dataset

Our dataset is a subset of a larger dataset available on Kaggle Kaggle dataset link. The original dataset was divided into 50 files for easier access and download. Due to its large size, we have randomly selected a subset, creating a dataset of 600 videos. We maintained a 2:1 ratio of fake to real videos, meaning there are two fake videos for every real video. Details of the real and fake videos, along with their respective tags, are stored in a JSON file named `merged_output_selected_videos_original_format.json`. This JSON file contains information for each video, including its name, label ("Real/Fake"), and, if the video is fake, the name of the original video.

Data available at: Video dataset link

1.2 Collection of Reviews

We have hosted a website on the IITK cloud website link [1]. The site features a signup option for first-time users (shown in Figure 1) and a log in option for returning users (shown in Figure 2). During the initial signup, we collect the user's email ID, gender, age, department, and branch, which are used in our analysis. After signing up, users can log in and start reviewing the videos.

User Review dataset link

2 Methodology

We have developed a website hosted on the cloud to gather reviews from users of various backgrounds. Our dataset is also uploaded to the server. We have collected reviews from approximately 50 users, totalling around 1500 reviews. Each review includes the following user details: name, email, age, gender, and department. Each user has provided feedback on the videos, indicating whether they are real or fake, along with a remark. The Github link to the website code can be accessed at Github Link.

2.1 User Reviews

2.1.1 Steps to Take User Reviews

Step 1: User Signup

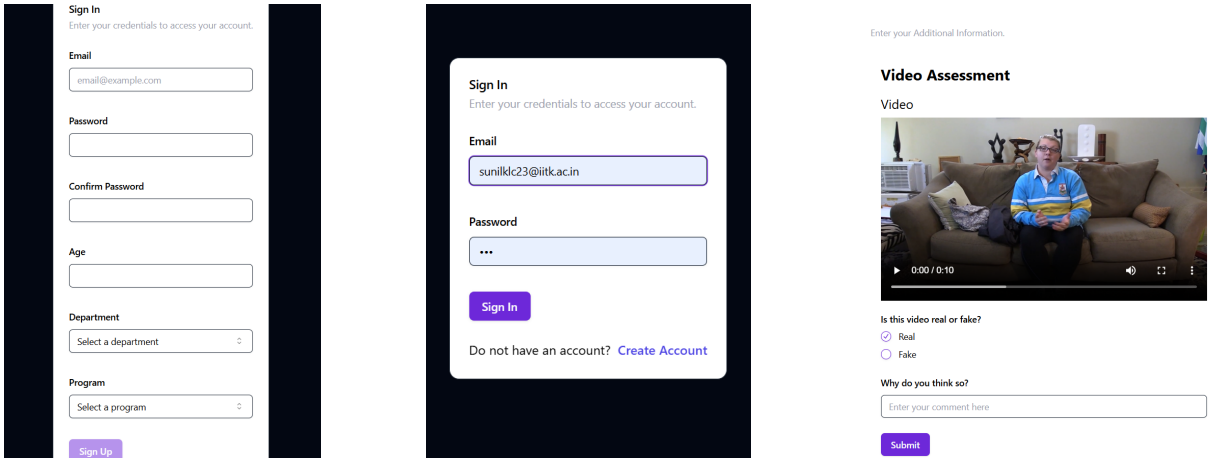
First, users are prompted to sign up on the website. The signup process requires them to provide their basic information such as email, gender, age, department, and branch. This data will be used later for analysis.

Step 2: User Sign-in

Once the user has successfully signed up, they need to sign in to the website. This allows users to access their accounts securely and start interacting with the platform.

Step 3: Reviewing Videos

After signing in, users can begin reviewing the videos on the platform. The reviews are collected for analysis, and users can rate the content based on their experience.



(a) Signup Page (b) Sign-in Page (c) Video Review Page

Figure 1: Steps for User Review Process

While taking feedback we have observed that some of the users have registered themselves but didn't give any feedback so we have removed all the users who have not given any feedback.

2.2 Demographic Analysis:

2.2.1 Average Response per Person:

Each user provided an average of 30 reviews, contributing to a total of approximately 1500 reviews. This metric helps us understand the engagement level of each participant and the overall comprehensiveness of the feedback collected.

2.2.2 User Details and Background:

We gathered detailed information about each user, including their name, email, age, gender, and department. This comprehensive data collection ensures that we can analyze feedback about diverse user demographics.

2.2.3 Age-wise Distinction:

The data was categorized based on age groups to identify patterns and trends in the feedback provided by users of different ages. This analysis helps us understand how perceptions and opinions may vary across various age demographics.

2.2.4 Branch-wise Distinction:

Users were also categorized based on their branch or field of study/work. This distinction allows us to observe feedback trends within different academic or professional branches, providing insights into how specific expertise or interests might influence user reviews.

2.2.5 Department-wise Distinction:

Similarly, categorizing users by their departments enabled us to analyze the feedback within different organizational or academic departments. This helps in understanding department-specific trends and any unique perspectives that may arise from users' departmental affiliations.

2.3 Methods to analyse the reviews

Further, we have analysed the top 5 videos from each category (Real/Fake) which are correctly identified.

2.4 Confidence-Level-Analysis

2.4.1 TP, FP, FN, TN for Real and Fake Videos :

In the context of video classification as real or fake, the terms True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) represent the results of a classification model's performance:

- **True Positive (TP)**: This occurs when the user correctly identifies a **real video** as **real**.
- **False Positive (FP)**: This occurs when the user incorrectly classifies a **real video** as **real**.
- **False Negative (FN)**: This happens when the user incorrectly classifies a **fake video** as **real**. The video was fake, but the user labeled it as real.
- **True Negative (TN)**: This occurs when the model correctly identifies a **real video** as **real**. The video was indeed real, and the model classified it correctly as real.

For Real Videos:

- **TP (True Positive)**: The real video that was correctly identified as real.
- **FP (False Positive)**: The real video that was incorrectly identified as fake.

For Fake Videos:

- **TN (True Negative)**: The fake video that was correctly identified as fake.
- **FN (False Negative)**: The fake video that was incorrectly identified as real.

2.5 Confusion Matrix for Real and Fake Videos

A **Confusion Matrix** is a table used to describe the performance of a classification model. For binary classification (real or fake), the confusion matrix is typically structured as follows:

	Predicted Real	Predicted Fake
Actual Real	TP	FP
Actual Fake	FN	TN

Where:

- **TP** is the number of real videos correctly identified as real.
- **FP** is the number of real videos incorrectly identified as fake.
- **FN** is the number of fake videos incorrectly identified as real.
- **TN** is the number of fake videos correctly identified as fake.

Interpretation:

- **Accuracy**: The proportion of correct predictions (both TP and TN) to the total number of videos:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

- **Precision (for Fake videos)**: The proportion of correct fake video predictions to all predictions made as fake:

$$\text{Precision for Fake} = \frac{TN}{TN + FP}$$

- **Recall (for Fake videos)**: The proportion of actual fake videos that were correctly identified as fake:

$$\text{Recall for Fake} = \frac{TN}{TN + FN}$$

- **F1-Score (for Fake videos)**: The harmonic mean of precision and recall for fake videos:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

These metrics help evaluate the model's performance comprehensively, taking both true and false classifications into account.

2.6 ROC Curve

The Receiver Operating Characteristic (ROC) Curve is a graphical representation of a classification model's performance. It illustrates the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) for different classification thresholds. The ROC curve helps evaluate the overall ability of the classifier to distinguish between classes, especially when dealing with imbalanced datasets.

The two key components in the ROC curve are:

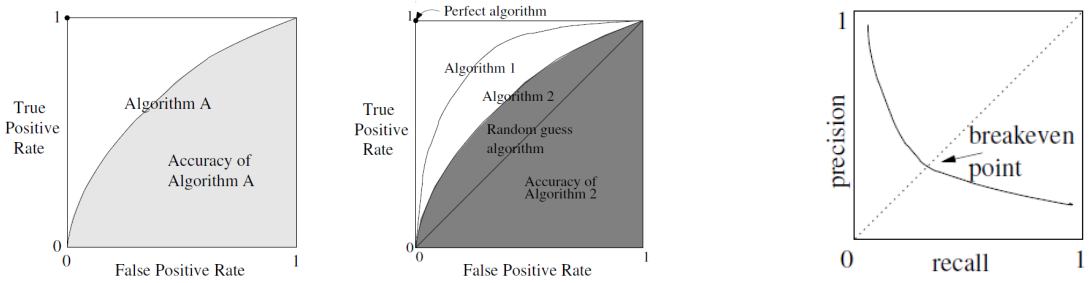
- **True Positive Rate (TPR)** or *Sensitivity or Recall*:

$$TPR = \frac{TP}{TP + FN}$$

- **False Positive Rate (FPR)**:

$$FPR = \frac{FP}{FP + TN}$$

The ROC curve is created by plotting the TPR on the y-axis and the FPR on the x-axis Fig. 2(a), with the threshold varying from 0 to 1. A model with a higher TPR and lower FPR will yield a curve closer to the upper left corner, which signifies better performance. The area under the curve (AUC) is a metric that summarizes the performance, with higher values indicating better classification performance. The AUC value ranges from 0 to 1, where 1 indicates perfect classification and 0.5 indicates a random classifier.



(a) (b)

Figure 2: ROC and PR Curves

2.7 Precision-Recall Curve

The Precision-Recall (PR) Curve is another graphical tool used to evaluate the performance of a classification model, especially when dealing with imbalanced classes. Unlike the ROC curve, which plots True Positive Rate (TPR) against False Positive Rate (FPR), the PR curve plots Precision versus Recall for different threshold values Fig. 2(b).

- **Precision** (also called Positive Predictive Value):

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall** (also called Sensitivity or True Positive Rate):

$$\text{Recall} = \frac{TP}{TP + FN}$$

The PR curve helps in understanding the trade-off between precision and recall. A model with a high precision and high recall indicates that it is classifying most of the positive samples correctly, and not misclassifying too many real samples as fake. The PR curve is particularly useful when the positive class (e.g., fake videos) is of greater interest, as it directly measures the performance of the positive class.

A high area under the PR curve (AUC-PR) signifies a better-performing model. Unlike the ROC curve, the PR curve is preferred when dealing with imbalanced datasets, as it focuses on the performance of the minority class.

2.8 Word Cloud Methodology

The word cloud methodology for identifying real and fake videos involves creating visual representations of the most frequently occurring words in the video metadata, comments, and descriptions. For both real and fake videos, word clouds are generated to highlight the prominent words and phrases that appear. This process involves several steps: collecting and preprocessing the text data, removing common stop words, and applying text normalization techniques. By comparing the word clouds of real and fake videos, distinct patterns and keywords can be identified, providing insights into the linguistic characteristics that differentiate authentic content from deceptive content. This analysis aids in understanding the contextual and semantic differences, which can be further utilized in developing automated detection algorithms.

2.9 TF-IDF: Term Frequency - Inverse Document Frequency

The TF-IDF (Term Frequency - Inverse Document Frequency) model is a widely used method in information retrieval and text mining for evaluating the importance of a word (or term) within a document relative to a corpus of documents. It takes into account both the frequency of a term within a document and how commonly it appears across all documents in the corpus. Below is an explanation of the components and the formula of TF-IDF.

2.9.1 Term Frequency (TF)

Given a corpus of documents D and a keyword t, the Term Frequency (TF) measures the importance of the keyword t within a particular document $d \in D$. If a keyword t occurs many times in a document d, it is likely more important for that document.

The formula for term frequency (TF) is as follows:

$$\text{tf}(t, d) = \frac{\text{Number of occurrences of } t \text{ in } d}{\text{Total number of terms in } d}$$

This gives us the proportion of times the term t appears in document d. It captures how frequently a term occurs in a document, with higher frequencies indicating that the term is important to the document.

Explanation: The term frequency is a basic metric that simply counts how often a term appears in a document. However, long documents may naturally have higher term counts, so TF is sometimes normalized by dividing by the total number of terms in the document to account for document length.

2.9.2 Inverse Document Frequency (IDF)

While term frequency measures how common a word is within a particular document, it does not account for how common or rare the word is across the entire corpus of documents. A term that occurs frequently in a single document may not be particularly useful if it is very common across the whole corpus. To address this, we introduce Inverse Document Frequency (IDF).

The IDF of a term t is a measure of how important the term is within the entire corpus D. It is computed by looking at the number of documents in which the term t appears.

The formula for inverse document frequency (IDF) is:

$$\text{idf}(t, D) = \log\left(\frac{|D|}{|\{d : t \in d\}|}\right)$$

Where:

- $|D|$ is the total number of documents in the corpus.
- $|\{d : t \in d\}|$ is the number of documents in which the term t appears.

Explanation: If the term t appears in many documents, its IDF will be low because it does not distinguish any particular document from others. If the term appears in fewer documents, its IDF will be higher, indicating that it is rare and thus more important for differentiating documents.

2.9.3 TF-IDF: Combining TF and IDF

Together, TF and IDF provide a measure of how important a term is in a specific document relative to the corpus. The TF-IDF score for a term t in document d is the product of the term frequency (TF) and the inverse document frequency (IDF):

$$\text{tf-idf}(t, d) = \text{tf}(t, d) \times \text{idf}(t, D)$$

Explanation: The TF-IDF score combines the information from both the term's frequency within a document (TF) and how rare it is across the corpus (IDF). The result is a weighted measure of the term's importance within the document in the context of the entire corpus. Terms with a high TF and low IDF will have a higher TF-IDF score, indicating they are important for that document but are not too common across the corpus.

2.9.4 Relevance in Our Analysis

In our analysis, TF-IDF played a pivotal role in comparing response patterns for real and fake videos, uncovering distinctive terms and phrases, and identifying shared characteristics. The methodology involved:

- Response Representation: Each user's response was represented as a vector of TF-IDF scores, capturing the importance of terms and n-grams (e.g., bigrams and trigrams) within the response context.
- Analysis of Real and Fake Responses: By computing the TF-IDF scores separately for real and fake responses and removing terms common to both, we focused on unique patterns specific to each category.

- Clustering and Similarity: The TF-IDF representation enabled us to compute similarity metrics (cosine similarity) to better understand the user response variation for a particular video.
- Interpretation of Key Terms: The top terms with the highest TF-IDF scores were analyzed to gain insights into user reasoning for identifying real and fake videos.

2.9.5 Practical Utility of TF-IDF

The TF-IDF methodology facilitated a deeper understanding of user behaviour by highlighting distinctive linguistic patterns:

- Discriminative Features: The combination of TF and IDF provided a robust mechanism to identify terms that are highly indicative of real or fake videos.
- User-Level Insights: By aggregating TF-IDF scores across responses, we analyzed individual-level response tendencies, aiding in the identification of high and low-performing users.

2.10 Cosine Similarity

Cosine similarity is a metric used to measure the similarity between two non-zero vectors in a multidimensional space. It calculates the cosine of the angle between two vectors, ranging from -1 to 1. In the context of text analysis, it is often used to compare document vectors created using methods like TF-IDF, where higher cosine similarity values indicate greater textual similarity.

The cosine similarity between two vectors A and B is defined as:

$$\text{Cosine Similarity}(A, B) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

Where:

- $\mathbf{A} \cdot \mathbf{B}$ is the dot product of vectors A and B.
- $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ are the magnitudes (or norms) of vectors A and B, respectively.

Cosine similarity focuses on the orientation of the vectors rather than their magnitude, making it especially useful for text data where the magnitude may not have meaningful interpretations.

2.10.1 Application of Cosine Similarity in the Project

In the project, cosine similarity was used to compare user responses to videos. The analysis aimed to identify patterns and similarities in reasoning among users for both real and fake videos. The specific steps involved were:

1. **Selection of Videos**:
   - The real and fake videos with the highest number of user responses were identified.
   - This ensured the analysis focused on videos with sufficient user input for meaningful comparisons.
2. **Representation of Responses**:

- User responses for the selected videos were converted into TF-IDF vectors. This representation captures the importance of terms within the response context, ensuring a consistent numerical format for similarity calculation.

3. **Calculation of Cosine Similarity**:

- For each pair of responses, cosine similarity was computed using the TF-IDF vectors.
- The resulting similarity values formed a cosine similarity matrix, where each entry represents the similarity between a pair of user responses.

4. **Visualization**:

- The cosine similarity matrices for the real and fake videos were visualized as heatmaps. These heatmaps highlighted clusters of similar responses and provided insights into the alignment of user reasoning across the dataset.

2.11 Temporal Analysis

Temporal analysis is a method used to study user performance trends over time. By examining the progression of user responses, it becomes possible to identify patterns and assess the evolution of accuracy on an individual or aggregate level. This analysis incorporates cumulative accuracy calculations and visualizations for the best and worst performers, as well as overall user performance.

2.11.1 Methodology

1. **Data Preparation**:

- Merged user responses with video ground truth labels using video ID as the key.
- Converted response timestamps to datetime format and sorted entries by user ID and time for chronological analysis.

2. **Accuracy Calculation**:

- Calculated response accuracy by comparing user predictions (`is_real_user`) with ground truth labels (`is_real_ground_truth`).
- Focused on active participants by excluding users with no responses.

3. **Identifying Best and Worst Performers**:

- Computed each user's overall accuracy as the mean of their individual response accuracies.
- Defined the best performer as the user with the highest average accuracy, and the worst performer as the one with the lowest among active users.

4. **Tracking Cumulative Accuracy Over Time**:

- Calculated cumulative accuracy per user by taking the running mean of their response accuracies, tracking progression over time.

5. **Visualization**:

- **Individual User Performance**:
  - Plotted cumulative accuracy over time for a high-response user.
  - Generated separate plots for best and worst performers to illustrate their respective temporal trends.
- **Aggregate Performance**:
  - Plotted average accuracy over time for all users to identify overall performance trends.

2.11.2 Justification for Temporal Analysis

Temporal analysis provides insights into both individual and collective performance patterns, enabling a deeper understanding of user behaviour over time:

- **Individual Trends**:
  - By analyzing cumulative accuracy over time, we can identify learning trends, consistency, and fluctuations in user performance.
  - Examining best and worst performers sheds light on factors that may influence accuracy, such as changes in viewing conditions (e.g., zooming in on videos causing blur, leading to misclassification of real videos as fake).
- **Aggregate Trends**:
  - Average accuracy plots reveal overarching patterns, such as whether users improve collectively over time or face challenges at specific moments.
- **Actionable Insights**:
  - Temporal trends help pinpoint critical moments where user engagement or support may be necessary, enabling targeted interventions.
  - Understanding how individual behaviours evolve provides context for designing strategies to improve overall accuracy and user experience.

By tracking and visualizing temporal trends, this analysis offers valuable perspectives on user behavior, highlighting patterns and anomalies that inform data-driven improvements and engagement strategies.

Overall, the temporal analysis provides a nuanced view of performance patterns, helping interpret user behaviour and evaluate the effectiveness of intervention strategies.

2.12 Automated Thematic Analysis

Thematic analysis is a qualitative research method for identifying and interpreting patterns (themes) within text data, such as interview transcripts or survey responses. It involves familiarizing oneself with the data, coding important segments, grouping similar codes into themes, refining these themes, and then summarizing them to capture key insights. This method is valuable in exploring new research areas or understanding

diverse perspectives, as it highlights commonalities and differences within the data. In deepfake studies, thematic analysis can help reveal factors users consider when assessing video authenticity (e.g., body language or audio quality), providing insights to improve deepfake detection tools or educational efforts. Automated clustering can aid thematic analysis by grouping similar responses, offering a preliminary look at recurring themes in large datasets.

3 Results

This section presents the results of the video review data analysis. The data comprises a total of 600 videos, out of which 400 are real and 200 are fake. The study involved 48 students from IIT Kanpur, who provided a total of 1420 reviews. On average, each participant submitted 30 reviews.

3.1 Demographics of Participants

We have categorized the participants based on their age group, gender, department, and branch. Below Fig. 3is a breakdown of the participant demographics:

```
Total Number of Videos: 600
Total Number of Real Videos: 200
Total Number of Fake Videos: 400
------------------------------
Number of reviews: 1420
------------------------------
Number of Participants: 48
------------------------------
Average Response per user: 30
------------------------------
Age: 20-24, Total Participants 28
Age: 25-29, Total Participants 10
Age: 30-34, Total Participants 6
Age: 35-39, Total Participants 2
Age: 40+, Total Participants 2


Gender: F, Total Participants 7
Gender: M, Total Participants 39


Branch: B.Tech, Total Participants 5
Branch: M.Sc. (2 yr), Total Participants 1
Branch: M.Tech, Total Participants 35
Branch: Others, Total Participants 3
Branch: Ph.D, Total Participants 4


Department: Cognitive Science, Total Participants 1
Department: Computer Science and Engineering, Total Participants 33
Department: Electrical Engineering, Total Participants 1
Department: Materials Science Programme, Total Participants 1
Department: Materials Science and Engineering, Total Participants 1
Department: Mechanical Engineering, Total Participants 5
Department: Others, Total Participants 5
Department: Statistics, Total Participants 1
```

Figure 3: Demographics of Participants: Age, Gender, Branch, and Department

3.2 Confusion Matrix

The confusion matrix shown in Fig. 4 is related to the detection of real and fake user responses. Here's a breakdown of the information presented:

- **User Response: Real**
  - True Positive (Actual: Real, Predicted: Real) = 393
  - False Positive (Actual: Fake, Predicted: Real) = 450
- **User Response: Fake**
  - True Negative (Actual: Fake, Predicted: Fake) = 492
  - False Negative (Actual: Real, Predicted: Fake) = 85

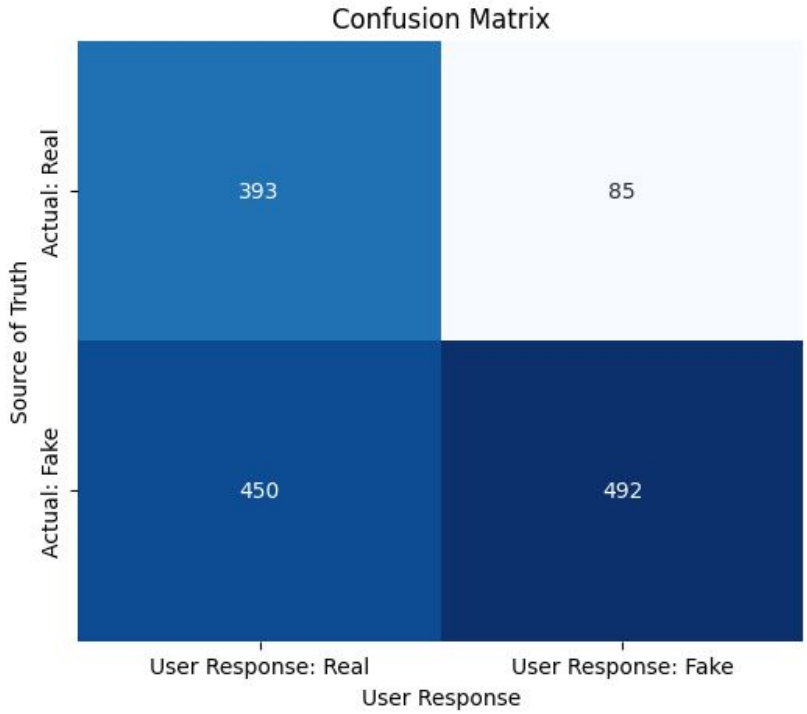


Figure 4: Confusion Matrix for Overall Data

3.3 Error Parameters

The error parameters of the overall data, including precision, recall, and F1 score etc, are shown as in below figure Fig. 5.

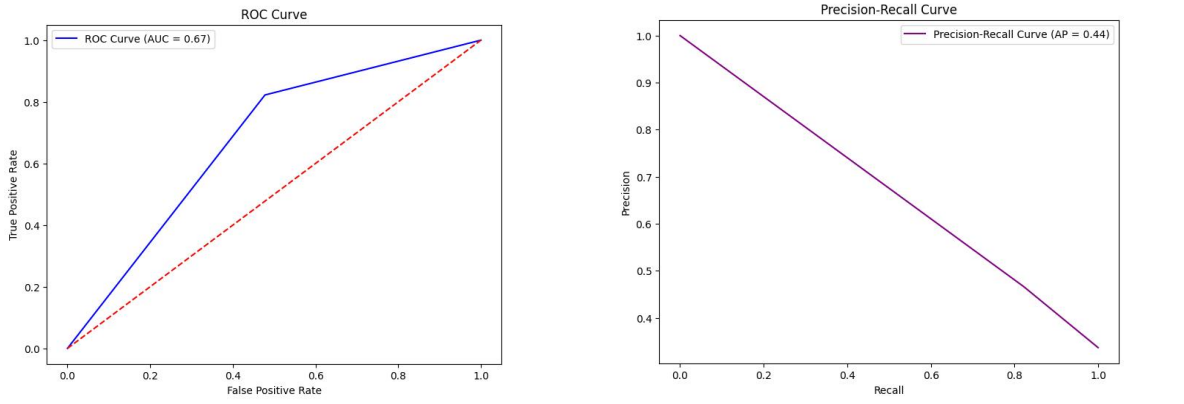
	Metric	Value
0	Accuracy	0.6232394366197183
1	Precision (REAL)	0.46619217081850534
2	True Negative Accuracy (FAKE)	0.8526863084922011
3	Recall	0.8221757322175732
4	True Negative Rate (Recall Fake)	0.5222929936305732
5	False Positive Rate	0.47770700636942676
6	False Negative Rate	0.17782426778242677
7	Error Rate	0.3767605633802817
8	F1 Score	0.595003785011355

ROC Curve

Figure 5: Confusion Matrix for Overall Data

3.4 ROC Curve

The ROC (Receiver Operating Characteristic) curve in Fig. 6(a) shows the trade-off between the true positive rate (TPR) and the false positive rate (FPR) for different classification thresholds. The area under the ROC curve (AUC) is 0.67, indicating a moderately good performance in discriminating real from fake videos. The ROC curve starts from the bottom-left corner, where the TPR and FPR are both 0, and curves towards the top-left corner, where the TPR is 1 and the FPR is 0. An ideal classifier would have a ROC curve that goes straight up to the top-left corner, maximizing the TPR while minimizing the FPR. The ROC curve in the image shows a steady increase in the TPR as the FPR increases, suggesting the users have reasonable ability to distinguish real videos from fake ones, but there is still room for improvement in the classification performance.



(a) ROC (b) PR

Figure 6: Results: (a) ROC curve, (b) PR curve

3.5 PR Curve

This graph in Fig. 7(b) shows the precision-recall curve for a model with an average precision (AP) of 0.44. The x-axis represents the recall, which is the proportion of true positives that are correctly identified. The y-axis shows the precision, which is the proportion of true positives among all the positive predictions. The

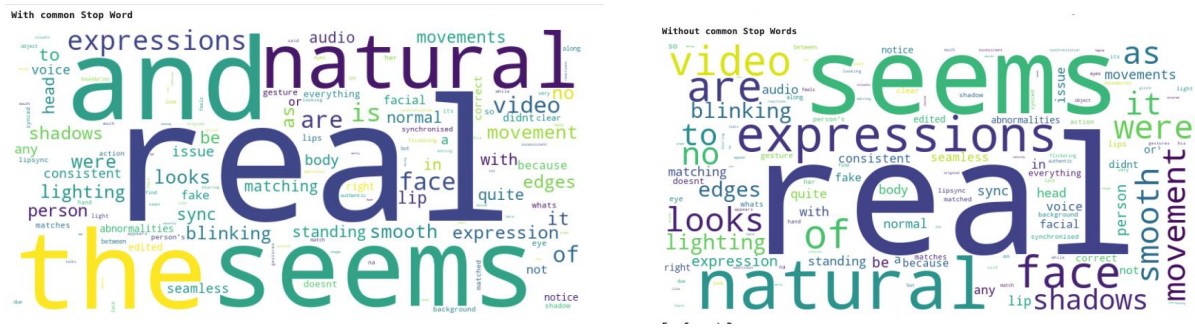
curve starts at a high precision around 0.9 and gradually decreases as the recall increases. This indicates that the model can achieve high precision in identifying real videos, but as it tries to capture more of the true positives (real videos), the precision starts to drop. The area under the curve (AUC) is 0.44, which suggests the model has a moderate ability to distinguish real from fake videos.

3.6 Wordnets

The following figures illustrate the wordnets created for the analysis of real and fake videos.

3.6.1 Wordnets for Real Videos

The wordnets for real videos show the distribution of words used in the reviews of real videos. We have generated wordnets with and without stop words to better understand the impact of common words on the analysis.

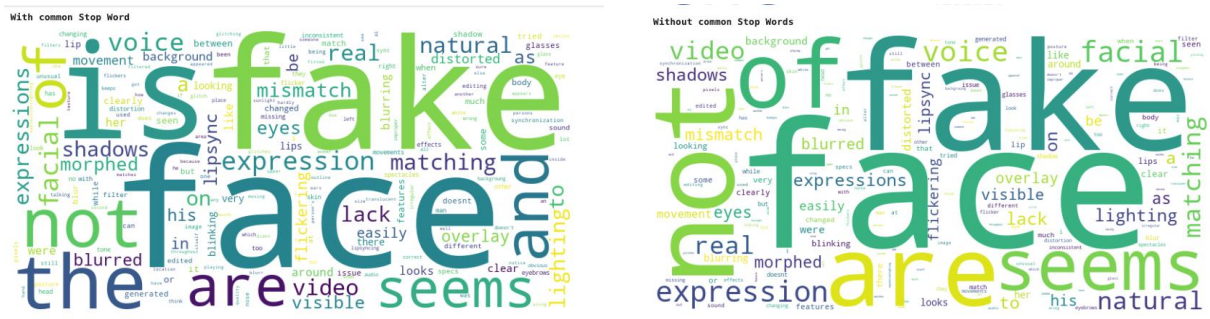


(a) With Stop Words (b) Without Stop Words

Figure 7: Wordnet for Real Videos: (a) With Stop Words, (b) Without Stop Words

3.6.2 Wordnets for Fake Videos

Similarly, the wordnets for fake videos highlight the vocabulary used in the reviews of fake videos. These visualizations help in understanding the linguistic differences between real and fake video reviews.



(a) With Stop Words (b) Without Stop Words

Figure 8: Wordnet for Fake Videos: (a) With Stop Words, (b) Without Stop Words

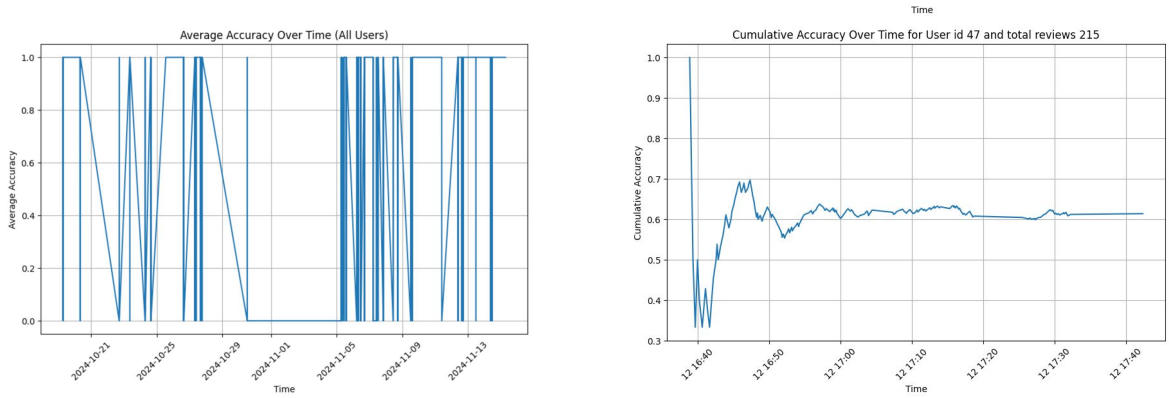
3.7 Temporal Analysis

Overall Temporal Analysis

The graph in Fig. 9(a) illustrates the average accuracy over time for all users of the system, highlighting key trends and fluctuations in performance. The accuracy varies significantly throughout the time period, ranging from approximately 0.2 to over 0.9. Several sharp peaks and valleys are observed, indicating substantial changes in performance at different points. However, the overall trend reveals a general improvement in accuracy, starting from a low of around 0.2 and reaching a high of nearly 0.9 towards the end. The most recent data points consistently show accuracy levels above 0.8, suggesting that the system is performing well in the latest time period. Overall, the graph provides a comprehensive view of the system's accuracy, showcasing both the volatility and the steady improvement in performance over time.

Most Performer:

The graph in Fig. 9(b) illustrates the cumulative accuracy over time for a specific user (user ID 47) along with the total number of reviews (215). Key observations include a high starting accuracy around 0.9, followed by significant fluctuations over time, with the accuracy dropping to approximately 0.5 at certain points. Several sharp peaks and valleys are evident, indicating considerable changes in performance throughout the observed period. Overall, the trend shows the cumulative accuracy remaining within the 0.5 to 0.8 range, without a clear upward or downward trajectory. The most recent data point places the cumulative accuracy at around 0.65.



(a) Overall

(b) Most Performer

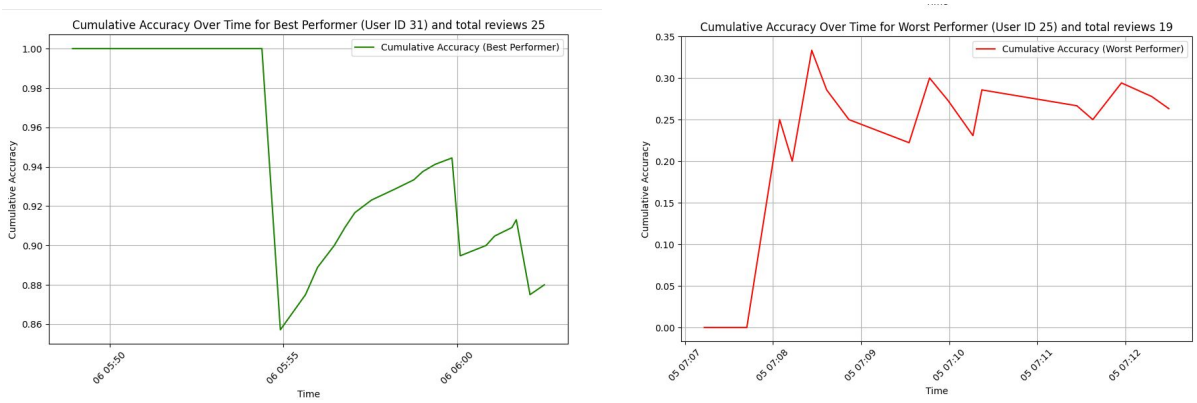
Figure 9: Temporal Analysis

Best Performer:

The graph Fig. 10(a) illustrates the cumulative accuracy over time for the best performing user (User ID 31), who has a total of 25 reviews. Initially, the accuracy starts at approximately 0.96 and shows a steady increase over time, ultimately reaching nearly 0.99 by the end. This graph highlights the user's consistent and strong performance.

Worst Performer:

The graph Fig. 10(b) depicts the cumulative accuracy over time for the worst performing user (User ID 25), who has a total of 19 reviews. The accuracy begins at approximately 0.32 and fluctuates significantly, dropping to as low as around 0.15 at one point. Overall, this user's performance is considerably weaker compared to the best performer, with the cumulative accuracy staying below 0.35 for most of the observed period.



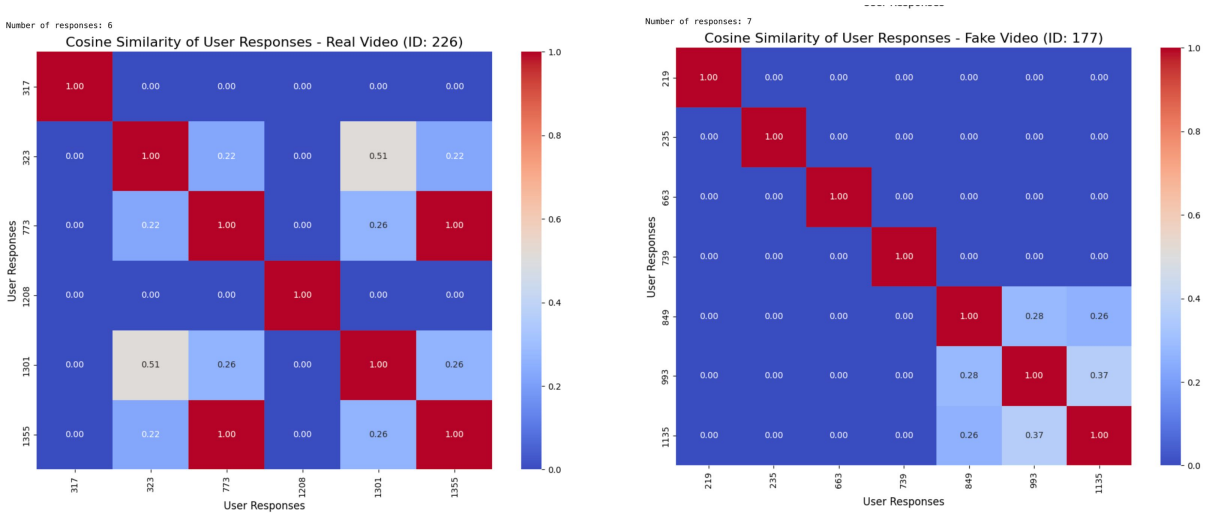
(a) Best (b) Worst

Figure 10: Temporal Analysis

3.8 Cosine Similarity (Fake and Real)

The following results Fig. 11(a) shows the cosine similarity matrix between a set of user responses for a real video. Both the x-axis and the y-axis represent different user responses and each cell in the matrix shows the cosine similarity score between two specific responses.

The image reveals that the cosine similarity values vary significantly across the different user responses. For example, responses 1355 and 773 have a cosine similarity of 1.0, indicating a high level of similarity. However, the responses 323 and 773 have much lower cosine similarity values, around 0.22. This suggests that the user responses are not consistently similar to each other, and there is a range of response quality or relevance.



(a) Real (b) Fake

Figure 11: Cosine Analysis

The image Fig. 11(b) shows the cosine similarity between a set of user responses for a fake video. The off-diagonal elements are zero which suggests that the responses are largely dissimilar. This implies that users described their observations or reasoning differently, with minimal overlap in wording or content. A few response pairs, such as responses 820 and 964 (0.28 similarity) and 964 and 1104 (0.37 similarity), have moderate similarity scores, represented by light blue. This suggests that these responses share some common words or themes but still have substantial differences in phrasing or content. This plot suggests a high level of variability in user responses, with minimal commonality in language or structure.

3.9 Thematic Analysis

Fig. 12 shows a thematic analysis of words associated with real and fake videos, categorized into 5 different themes.

Theme 1 highlights words related to facial features and visual qualities, such as ”smooth”, ”expressions”, and ”distorted”. This suggests that users perceive distinctions in the visual characteristics of real vs. fake videos.

Theme 2 focuses on terms related to the authenticity of the video, with words like ”real”, ”fake”, ”facial”, and ”features”. This indicates that users identify cues about the underlying nature of the video as being real or artificially created.

Theme 3 covers words associated with the overall video properties, including ”video”, ”looks”, ”voice”, ”normal”, and ”person”. This suggests that users assess factors like the look, sound, and coherence of the video when distinguishing real from fake content.

Theme 4 deals with more nuanced perceptual qualities, such as ”natural”, ”expressions”, ”shadows”, ”lighting”, and ”mismatch”. Users may rely on these subtler visual and contextual cues to identify manipulated videos.

Theme 5 encompasses a range of words related to the previous themes. The repeated nature of the theme suggests that we are running out of the unique themes.

Overall, the thematic analysis suggests that users employ a diverse set of visual, auditory, and contextual

cues to differentiate real videos from manipulated or synthetic ones. The words within each theme provide insight into the specific perceptual factors that users consider when making this distinction.

```
Theme 1:
face | smooth | expressions | morphed | distorted | overlay | blurred | flickering | different | features

Theme 2:
real | yes | filters | faces | facial | fake | false | faxce | feature | features

Theme 3:
real | video | looks | voice | normal | eyes | lip | seamless | issue | person

Theme 4:
natural | expressions | shadows | lighting | lack | matching | blinking | lip | sync | mismatch

Theme 5:
fake | facial | expression | voice | looking | backgroung | face | look | expressions | feature
```

Figure 12: Thematic Analysis of Videos

3.10 TF-IDF Analysis

The graph 13 shows the top unique words that users have marked as ”real” in the context of the given video content. The words with the highest TF-IDF (term frequency-inverse document frequency) scores are displayed, indicating their importance and distinctiveness in the ”real” user responses. The top words include ”smooth”, ”normal”, ”person”, ”blinking”, ”movements”, ”seamless”, ”edges”, ”issue”, ”body”, ”audio”, ”quite”, ”head”, ”standing”, and ”consistent”. These words suggest that users perceive the ”real” video content to have smoothness, normalcy, clear personification, natural movements, seamless edges, and consistent audio quality, among other attributes.

The second part of graph 13 shows Unique Top Words (User Marked as Fake) top unique words that users have marked as ”fake” in the context of the given video content. Similar to the first image, the words with the highest TF-IDF scores are displayed, indicating their importance and distinctiveness in the ”fake” user responses. The top words include ”fake”, ”morphed”, ”blurred”, ”overlay”, ”flickering”, ”visible”, ”distorted”, ”lack”, ”mismatch”, ”background”, ”easily”, ”blurring”, and ”clear”. These words suggest that users perceive the ”fake” video content to have artificial or morphed elements, blurriness, overlays, flickering, lack of clarity, and other visual discrepancies that make the content appear less genuine or authentic.

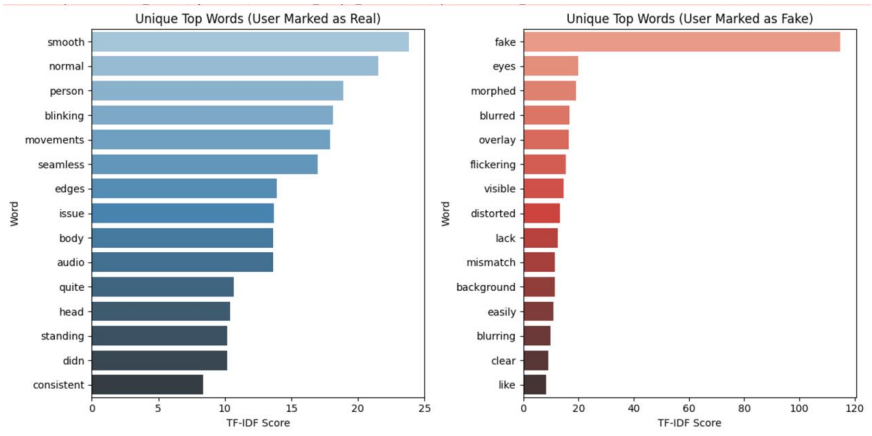


Figure 13: TF-IDF Analysis for Videos

Challenges Faced

1. **Networking Issues in Deployment:** During website deployment, unforeseen networking problems caused delays. These issues could be attributed to misconfigurations of firewalls and lack of internet access. Improved pre-deployment checks and better documentation of network requirements can mitigate such problems in future projects.

2. **Dataset Segregation:** Segregating videos from a large dataset proved cumbersome, especially with varied file formats and metadata inconsistencies. Automating this process with scripts significantly reduces manual effort.

3. **Quality of User Feedback:** Users often provided insufficient or unclear reasons for their judgments on whether videos were real or fake. This highlights the need for clearer guidelines, examples, or incentives to encourage meaningful feedback, ensuring higher-quality data for analysis.

4. **User Engagement with Reviews:** Getting users to review more than ten videos was challenging, likely due to fatigue or lack of interest.

By addressing these challenges through automation, user-friendly interfaces, and motivation strategies, the platform can be made more efficient and appealing for contributors.

4 Conclusions and Future Directions

4.1 Conclusion

- Users displayed moderate accuracy in detecting fake videos, achieving a score of 0.85. This suggests that users are relatively adept at identifying deceptive content.
- In contrast, the precision for real videos was significantly lower, at 0.45. This indicates:
  - A possible bias towards skepticism, leading users to erroneously classify real videos as fake.
  - Potential challenges in identifying markers of authenticity compared to markers of deception.
- Linguistic analysis of user feedback revealed clear differences in focus:
  - For fake videos, users predominantly centered their attention on facial features, such as unnatural expressions or distortions.
  - For real videos, decisions were driven by the overall naturalness and realism of the content.
- There was no statistically significant variation in performance across different demographic groups, suggesting that user behavior and perception were consistent irrespective of background.
- These findings underline key challenges in human-based detection of deepfakes:
  - The skepticism bias may undermine trust in legitimate content.
  - A lack of consistent visual or contextual markers for authenticity contributes to misclassification.

4.2 Future Works

1. **Integration with AI Models**: Develop advanced AI algorithms to compare user reviews with automated detection results, enhancing overall accuracy.
2. **Educational Tools**: Incorporate tutorials and interactive content to educate users on identifying deepfakes, fostering digital literacy.
3. **Real-Time Feedback**: Create tools for real-time flagging and reporting of suspected deepfakes, accelerating response times to misinformation campaigns.
4. **Multilingual Support:** Adapt the system for regional languages in India, ensuring accessibility for diverse user groups.

5 Team Contributions

1. **Shivin Agarwal (231110048):**
   (a) Website FrontEnd
   (b) Website BackEnd
   (c) Website Integration
   (d) Website Deployment
   (e) Server Management
   (f) Wrote scripts to Extraction, Transformation and Loading of review dataset
   (g) Got a portion of Reviews on the website
   (h) Worked on the following items of Results:
      i. Extraction, Transformation and Loading of Review DataSet
      ii. Word Cloud
      iii. Temporal Analysis
      iv. Accuracy Metrics and Confusion Matrix
2. **Sunil Kumar (231110025):**
   (a) Video Dataset Cleaning.
   (b) Video Dataset Processing.
   (c) Video Dataset Loading on to the server.
   (d) Wrote scripts for Extraction, Transformation and Loading of Video Dataset
   (e) Got a portion of Reviews on the website
   (f) Worked on the following items of Results:
      i. Demographics of Review Dataset
      ii. Age, Gender, Branch and Dept
      iii. Comparison Analysis on Different demographics of the Dataset
3. **Yuvraj Raghuvanshi (241110084):**

(a) Sanctioned our server from IITK CSE Dept.

(b) Website BackEnd

(c) DataBase modeling

(d) Got a portion of Reviews on the website

(e) Worked on the following items of Results:

i. TF-IDF

ii. Cosine Similarity

iii. Thematic Analysis

References

[1] Site Administrator. *Sign In Page*. http://172.27.96.188/signin. Accessed: 2024-11-14.

[2] News Checker. "Deepfake Ads of Nita Ambani, Gautam Adani Endorsing Shady Gaming Apps Go Viral". In: *NewsChecker* (2022). URL: https://newschecker.in/fact-check/deepfake-ads-of-nita-ambani-anant-ambani-yogi-adityanath-gautam-adani-endorsing-shady-gaming-app-go-viral/.

[3] CNN. "Hong Kong Deepfake Scam Targets CFOs Using AI". In: *CNN* (2024). URL: https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html.

[4] MIT Technology Review. "Deepfake revenge porn is on the rise. How should we stop it?" In: *MIT Technology Review* (2021). URL: https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/.

[5] Hindustan Times. "5 Indian Celebs Hit by Deepfake Videos". In: *Hindustan Times* (2021). URL: https://www.hindustantimes.com/web-stories/entertainment/5-indian-celebs-hit-by-deepfake-101701165142127.html.