Project Report
INFORMATION RETRIEVAL CSE 508

Yogesh Kaushik	Jahanvi Bakshi	Prateek Mishra
2020163	2020069	2020102

Shourya Singh	Siddharth	Ritik Kirar
2021422	2021424	2021413

PROBLEM FORMULATION -

The aim of our project is to develop an app/website that simplifies the process of reading privacy policies by providing a summarized version of crucial aspects that users need to know before granting access to their personal information. We aim to address the growing concern about digital privacy and protect users from potential data exploitation. Our algorithm will utilize keywords to identify critical points in the policy, and we will optimize it to ensure a safer digital experience for users. Additionally, we will generate a safety score for each application based on the data collected in our database and incorporate Play Store ratings and application reviews to suggest alternate apps with similar features. Going forward, we also aim to increase our data corpus and utilize web scraping in order for the app to be able to access privacy policies from the Internet and summarize along with providing a privacy score. Our primary goal is to establish trust with users by being transparent and seeking legitimacy through app testing agencies.

MOTIVATION

Understanding privacy policies in the digital era is crucial for safeguarding personal data and ensuring digital privacy. With the increasing use of mobile applications, users often unknowingly expose sensitive information due to complex and lengthy privacy policies. App-Police seeks to address this issue by simplifying these policies, making them more accessible and understandable to users. By enhancing user understanding of privacy policies, App-Police aims to empower individuals to make informed decisions about their digital privacy and data security.

LITERATURE REVIEW -

Privacy policies are documents of a complex nature that elucidate the methodology adopted by an application or website for collecting, using, and distributing user data. The policies, which are typically lengthy and riddled with legal jargon, demand considerable effort from users to comprehend the data being collected and how it is utilized. In recent years, growing concerns regarding user privacy and data protection have led to increased research on simplifying privacy policies and enhancing their comprehensibility for the average user.

Several studies have been conducted to analyze the readability of privacy policies. In a 2011 study by Aleecia M. McDonald and Lorrie Faith Cranor, the authors found that the average privacy policy on a website required a college undergraduate level of reading ability, thereby making it difficult for the average user to understand. In another study in 2017 by Hana Habib and Yunan Chen, the privacy policies of the top 20 apps on the Google Play Store were analyzed, and it was found that most of these policies were written at a level beyond the recommended 8th-grade level.

Several tools have been developed to assist users in understanding privacy policies. The Usable Privacy Policy Project (UPP), for instance, is a tool developed by Lorrie Faith Cranor and her team at Carnegie Mellon University that analyzes privacy policies and provides a summary of the key points in plain language. The
Privacy Assistant is a browser extension that provides users with a summary of the privacy policy of the website they are visiting.

Several companies have developed their own systems for rating the safety and privacy of apps. For example, Google Play Protect is a service provided by Google that scans apps for malware and other security issues before they are downloaded onto a user's device. Similarly, AV-TEST is an independent organization that tests and rates antivirus and security software.
Several app discovery platforms allow users to search for apps based on specific features or functionality. The Google Play Store and Apple's App Store, for instance, have a "related apps" section that suggests apps similar to the one a user is currently viewing. Third-party app discovery platforms such as AppBrain and AppCrawlr also enable users to search for apps based on specific features or functionality.

In conclusion, significant research and development efforts have been made around privacy policies and app safety ratings. While several tools have been developed to assist users in understanding privacy policies and evaluating the safety of apps, there is still a need for a comprehensive solution that combines these features and provides users with an easy-to-understand summary of the privacy policies of different apps on the Play Store.

NOVELTY

App-Police utilizes the T5-BASE model to generate accurate and meaningful summaries of mobile application privacy policies. By summarizing and creating a database of all apps on the App Store, it eliminates the need for users to upload URLs manually. This approach simplifies the process, enhancing user engagement and experience. With a focus on user-centric design, App-Police ensures that users can easily understand and engage with privacy policies, empowering them to make informed decisions about their digital privacy and data security.

METHODOLOGY

5.1 Data Collection

The data collection process for App-Police involves extracting privacy policy text from official documents associated with applications on the Google Play Store. This ensures that the analysis is based on the most up-to-date and accurate information provided by the app developers. Additionally, basic information regarding the permissions requested by the application upon installation is verified from the mobile device's settings and stored in our backend database for subsequent analysis.

5.2 Database Design

The backend database of App-Police is designed to store comprehensive information about the applications, including the permissions they request, their Play Store ratings, and the features they offer. This data serves as the foundation for the app's privacy policy analysis and personalized application recommendations. The database design is illustrated in the provided ER diagram, which showcases the relationships between the various entities and attributes.

5.3 Information Retrieval Techniques

5.3.1 Text Preprocessing

To prepare the privacy policy text for summarization and analysis, App-Police employs several text preprocessing techniques. These include tokenization (breaking down the text into individual words or phrases), lemmatization (reducing words to their base or dictionary form), and the removal of punctuation marks and stop words (common words that do not contribute to the overall meaning). These preprocessing steps help to normalize the text and improve the accuracy of the subsequent analysis.

5.3.2 Text Summarization

App-Police utilizes the pre-trained "Transformer" model T5-BASE, developed by Google, for text summarization. T5 (Text-to-Text Transfer Transformer) is a state-of-the-art language model

that has been trained on a vast corpus of text data and fine-tuned for various NLP tasks, including summarization. By leveraging this model, App-Police can generate concise and meaningful summaries of the privacy policies, highlighting the key points users need to be aware of.

5.3.3 Inverted Indexing

To enable efficient retrieval of relevant information, App-Police employs an inverted indexing technique. This involves creating an index that maps each word or phrase to the documents in which it appears. By using inverted indexing, the app can quickly locate and retrieve the specific parts of the privacy policies that mention keywords or phrases related to data collection, sharing, or other privacy concerns. This allows for faster and more accurate analysis of the privacy policies.

5.3.4 Score Calculations

To assess the privacy and safety of an app, App-Police utilizes a weighted sum approach in its baseline evaluation. The app is categorized into three groups: "secure," "unsafe," and "moderate," based on the presence and frequency of specific keywords related to user privacy. These keywords include terms such as "data collection," "third-party sharing," "advertising," and others that indicate potential privacy risks. The overall app score is determined by calculating the weighted sum of the occurrences of these keywords in the privacy policy summary. This score provides users with a quick and easy-to-understand assessment of the app's privacy practices.

5.3.5 App Recommendation

Our application incorporates a built-in recommendation system that suggests safer alternatives when a new app is installed. This system uses the Jaccard similarity metric to measure the similarity between the feature vectors of the newly installed app and those already in our database. These feature vectors consist of all distinct feature IDs found within the app. Each app in the database is assigned a score, categorizing them as safe or unsafe. Using this score and considering the similarity between the new app and those in the database, our system recommends apps that are not only similar but also safer. The lower the score assigned to an app, the safer it is deemed to be. Therefore, our system offers alternative recommendations that are safer than the currently installed app.

5.3.6 ChatBot

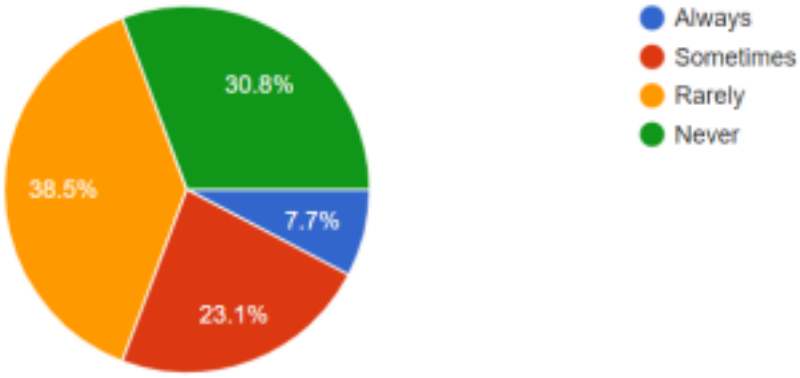
Our application includes a chatbot powered by the Gemini API, which has been optimized using data extracted from our privacy policy. This allows the chatbot to provide accurate and timely responses to user queries, showcasing our commitment to protecting their personal information. We converted our privacy policy data into JSON format to facilitate quick and accurate

response generation. By training the Gemini model on this preprocessed data, our chatbot now has a thorough understanding of our privacy policy guidelines.

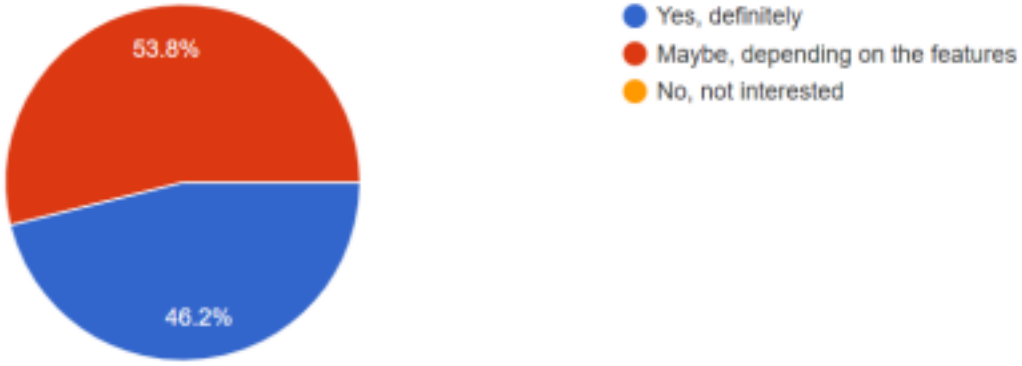
Apart from the above steps mentioned, we also conducted a survey to gain insights about how our Application would impact the users who will be using it.

These were some of the insights from the survey -

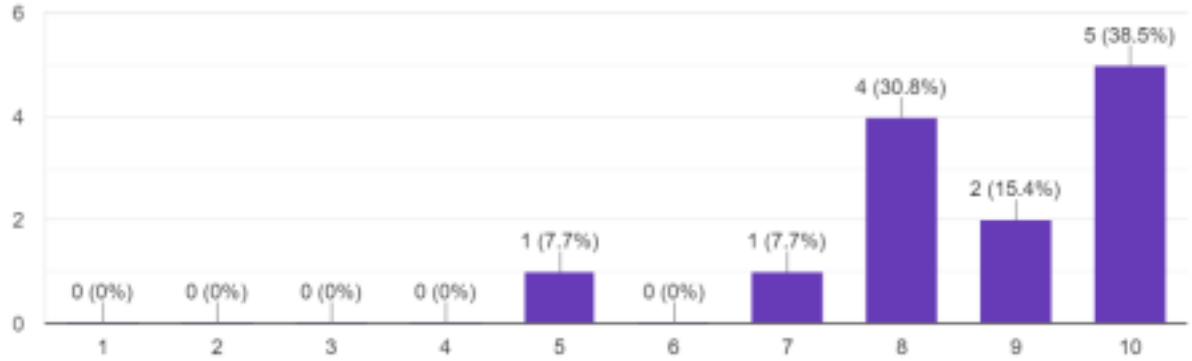
Q: How often do you read the privacy policies of the apps you download?



Q: Would you be interested in an app/website that summarizes the privacy policies of dif erent apps on the Play Store and provides a safety score for each app?



Q: How important (on a scale of 1-10) is the safety score of an app to you when downloading it?



Link to the form
https://docs.google.com/forms/d/e/1FAIpQLSddB0wtRKgm7AvSkYQVgYWDsY56c6JIbxZuj4rCVNGxYIEIMw/viewform?usp=sf_link-

DataBase
Our data was collected from the official privacy policy documents of the Google Play Store. Further, our database provides basic information about these applications, like their features, permissions, prices, etc.
The data is stored in our backend and can be found here for interpretation and analysis:

https://docs.google.com/spreadsheets/d/1QaVuMbXO3WaAVr8TNvhzkPBobxQkLNmW4EKUTUmbroY/edit#gid=1883706907

The layout of the database sufficed our needs; thus it has not been tampered with.

Database Tables -

1. Apps

App ID		App Name	Privacy Policy Summary Score Ratin	g	Paid

This is the main table that references TypeID from the Type Table.

2. Features

Feature ID Feature	

This is the table storing various features provided by applications.

3. Type

ID	Type

This is the table storing different types of applications.

4. Permission

Id	Permission

This is the table storing different accesses asked by applications.

5. App_join_permission

AppID

PermissionID

This is the table mapping various apps to the permissions they require. The appID references from the Apps table and permissionID references from the Permission table.

6. App_join_feature

AppID	Feature ID

This is the table mapping various apps to the permissions they require. The appID references from the Apps table and featureID references from the Feature table.

Code

Indexing

```
posDict['sharing']
```

✓ 0.0s

```
Output exceeds the size limit. Open the full output data in a text editor
{'Whatsapp': [768, 768],
 'WeChat': [1922, 1967, 2189, 2840, 1922, 1967, 2189, 2840],
 'Instagram': [953, 969, 1631, 1829, 1848, 953, 969, 1631, 1829, 1848],
 'Facebook': [977, 993, 1673, 1876, 1895, 977, 993, 1673, 1876, 1895],
 'Telegram': [817, 817],
 'Tumblr': [547,
  818,
  1753,
```

Query Processing

```
import query as pt
query = input("Enter the INPUT phrase query: ")
ans = pt.query_processing(posDict, query, AppID)
ans
```

✓ 5.6s

```
Query Using POSITIONAL INVERTED INDEX:
No of documents retrived:  5
```

```
def positional_query(query, posDict):

    potential = list(posDict[query[0]].keys())

    for i in range(1, len(query)):
        potential = AND(potential, list( posDict[query[i]].keys()) )

    ans = []

    for file in potential:
        temp = posDict[query[0]][file]
        for word in range(1, len(query)):
            temp = helper(temp, posDict[query[word]][file] )
        if( len(temp) != 0 ):
            ans.append(file)


    print("\nQuery Using POSITIONAL INVERTED INDEX: ")
    print("No of documents retrived: ", len(ans))

    return ans
```

Score calculation

```
def helper_function():
    query = ['data sharing', 'third party', 'advertisement', 'photo',
             'location', 'gallery', 'document', 'personal information']
    ans = []
    for word in query:
        temp = pt.query_processing(posDict, word, AppID)
        ans.append(temp)

    return score(ans, len(ans[0]))
```

Evaluation

In our evaluation, we utilized two metrics:

1. ROUGE (Recall-Oriented Understudy for Gisting Evaluation): ROUGE is a set of metrics that assesses the overlap between the generated summary and the reference summary based on n-gram matches. It calculates precision, recall, and F1 scores.

- OBTAINED ROUGE SCORE: 0.65

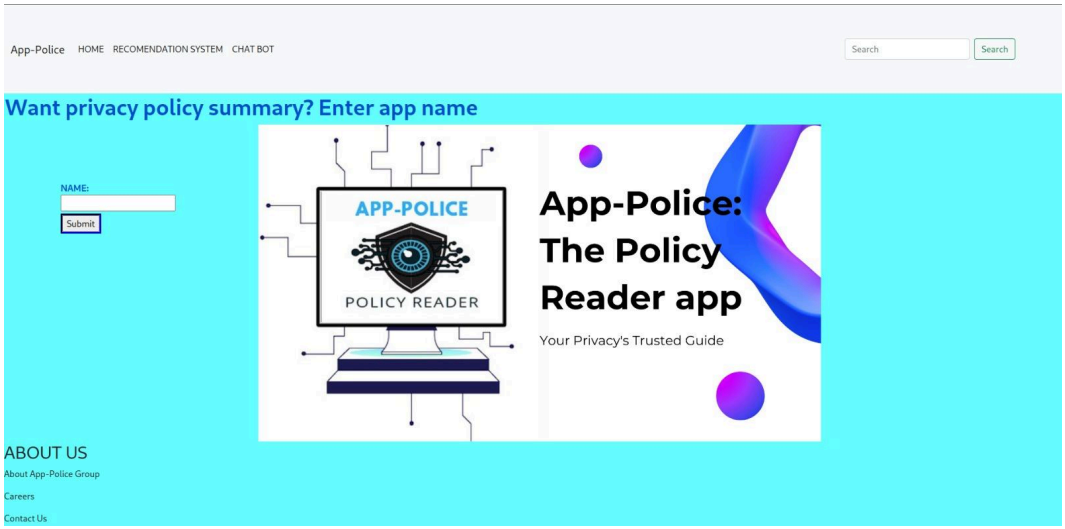
2. BLEU (Bilingual Evaluation Understudy): BLEU measures the similarity between the generated summary and the reference summary in terms of n-gram matches. It computes precision, recall, and the geometric mean of these two scores.

- OBTAINED BLEU SCORE: 0.7

NEW PRIVACY POLICIES

To handle the new privacy policy being added, we have used web scrapping, which will provide the privacy policy of the new applications.

Some Screenshots of our Frontend

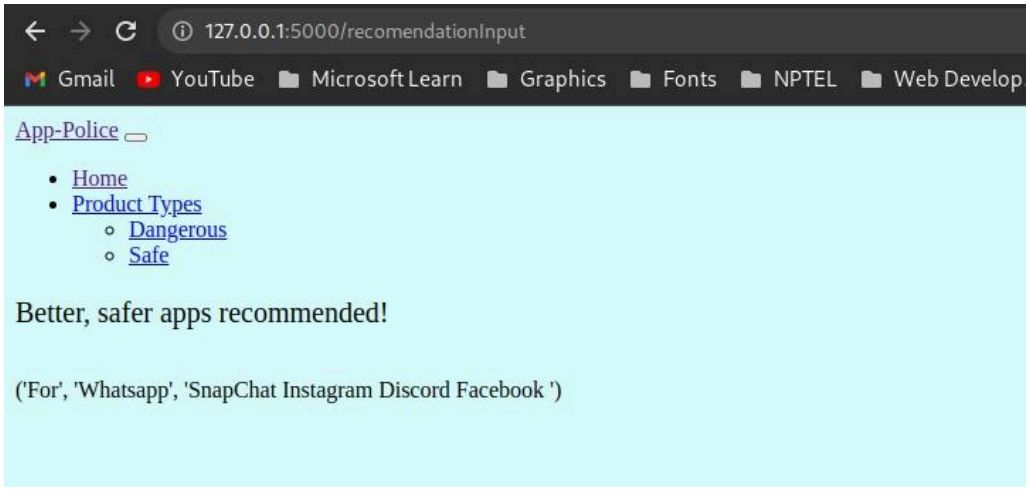


Summarization Results

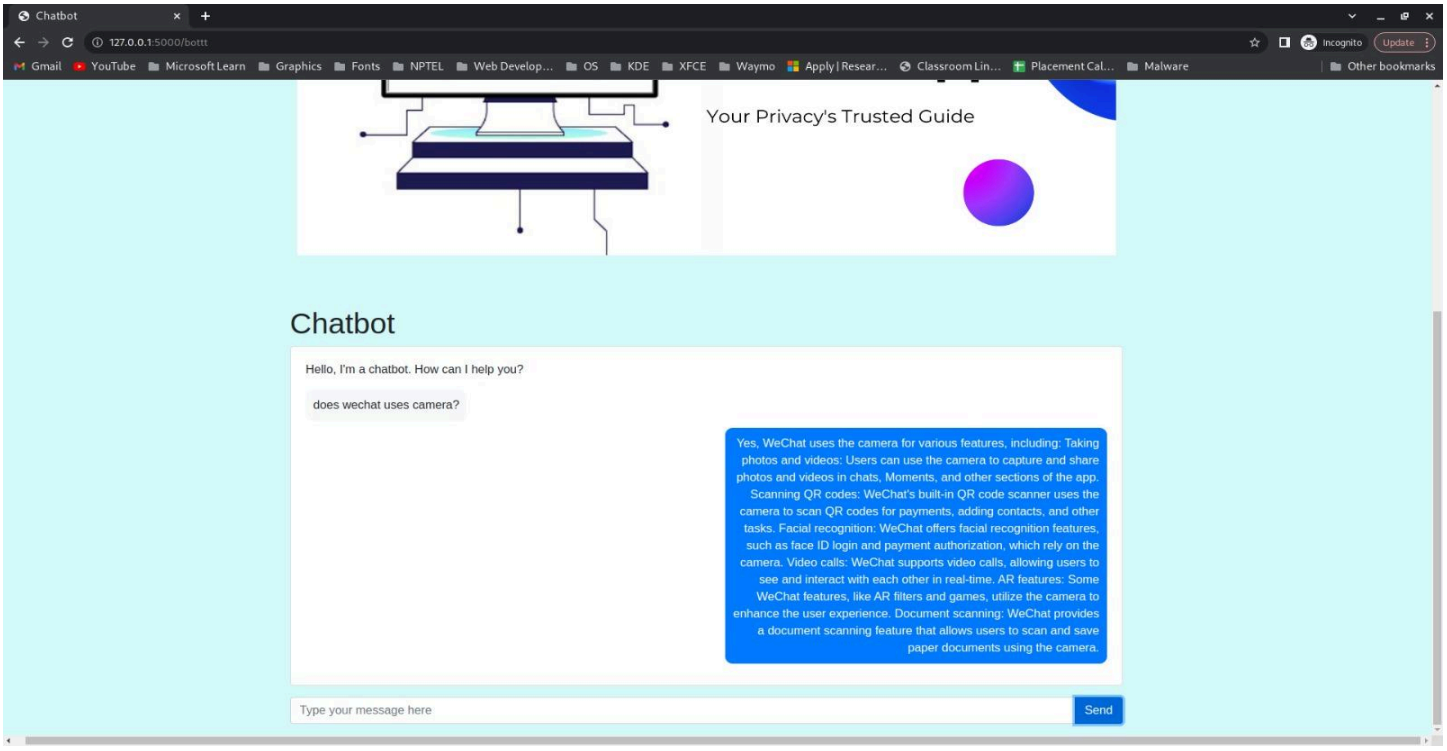
Search Results

App Name	Summary	Threat Score	Rating
Whatsapp	if you do not use our Services, we collect your information from your device. you can change your profile name, profile picture, and "about" information at any time, including when you choose to use the services, or send them to a third-party service provider or other Meta Companies products. We collect and use information we share with other companies, such as iCloud or Google Drive, to help us manage their communications with you.	6.153	4.000

App Recommendation



ChatBot



Github Repository Link: [Github repo]

Youtube Video Link : [Youtube Video]

CONTRIBUTORS-

1). Shourya Singh (2021422) - Front-end Development, Database Management, Indexing and NLP Techniques.

2). Ritik Kirar (2021413) - Front-end Development, Database Management, Indexing and Model Training.

3). Yogesh Kaushik(2020163) - Database

Management, Indexing, NLP Techniques and

Model Training.

4). Prateek Mishra (2020102) - Database

management, Indexing, NLP techniques, and Model

Training.

5). Siddharth (2021424) - Database

Management, Indexing, NLP techniques, and Model

Training.

6). Jahanvi Bakshi (2020069) - Database

Management, Indexing, and NLP techniques.