Baseline Results– Group 31 | Information Retrieval (CSE508) Winter 2024

# App-Police: An Application Policy Summariser

Jahanvi Bakshi

IIIT Delhi

[jahanvi20069@iiitd.ac.in](mailto:jahanvi20069@iiitd.ac.in)

Prateek Mishra

IIIT Delhi

[prateek20102@iiitd.ac.in](mailto:prateek20102@iiitd.ac.in)

Yogesh Kaushik

IIIT Delhi

[yogesh20163@iiitd.ac.in](mailto:yogesh20163@iiitd.ac.in)

Ritik Kirar

IIIT Delhi

[ritik21413@iiitd.ac.in](mailto:ritik21413@iiitd.ac.in)

Shourya Singh

IIIT Delhi

[shourya21422@iiitd.ac.in](mailto:shourya21422@iiitd.ac.in)

Siddharth

IIIT Delhi

[siddharth21424@iiitd.ac.in](mailto:siddharth21424@iiitd.ac.in)

## Problem Statement

We're working on a project to create an app that makes it easier for people to understand and digest privacy policies. These policies are often lengthy and complicated, leading many users to skip over them. Our app will provide a simplified version of these policies, highlighting the important points users need to know before they agree to share their personal information. Our main goal is to protect users from those who misuse their data without permission.

We understand that gaining users' trust and developing an algorithm that can distinguish between irrelevant and crucial parts of a privacy policy are big challenges. To tackle these issues, we plan to be transparent with our users and undergo testing with reputable agencies to establish legitimacy. Our algorithm will use keywords to identify key points in the policy, and we'll continuously refine it to ensure a safer digital experience for everyone.

Also, our aim is to generate a safety score for applications based on the data we collect in our database. Additionally, our app will suggest other applications that align with users' preferences while maintaining a good privacy   
score, depending on what features they're looking for.   
Moreover, it will integrate ratings from the app   
  
  
store and reviews of applications to enhance the recommendations.

## Literature Review

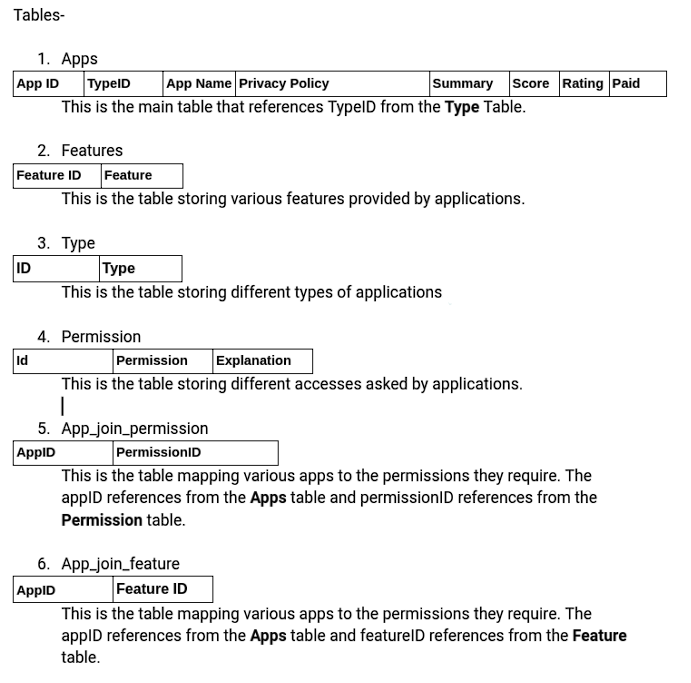
Numerous studies focus on summarizing and identifying unsafe apps through their privacy policies. For example, Narseo Vallina-Rodriguez et al.'s paper [1] "PrivacyGrade: Measuring the Privacy Behaviors of Smartphone Apps" introduces a methodology that evaluates smartphone apps' privacy behaviors by scrutinizing their privacy policies. The authors created PrivacyGrade, employing automated technology to analyze privacy policies and rate apps based on their privacy practices. In another study, Sushain Cherivirala et al.'s work [2] "Mining Privacy Policies for Better Mobile App Design" presents an approach utilizing natural language processing techniques to extract insights from privacy policies. They pinpoint standard privacy practices concerning data collection, sharing, and retention, which inform improved mobile app design.  
  
Moreover, in "Privacy Policy Analysis of Android VPN Apps" by Adwait Nadkarni et al. [3], the authors scrutinize the privacy policies of popular Android VPN apps and pinpoint issues related to data collection, sharing, and retention. They discovered that many VPN apps lacked clarity in disclosing their data practices, with some even gathering sensitive information without users' awareness or consent. Additionally, "PrivacyScore: Analyzing the Privacy Behaviors of Smartphone Apps at Scale" by Narseo Vallina-Rodriguez et al. [4] expands upon the PrivacyGrade approach to assess the privacy behaviors of a vast number of smartphone apps. Using PrivacyScore, the authors evaluated the privacy policies of over 5,000 Android apps and found that many apps inadequately disclosed their data practices or provided misleading information.

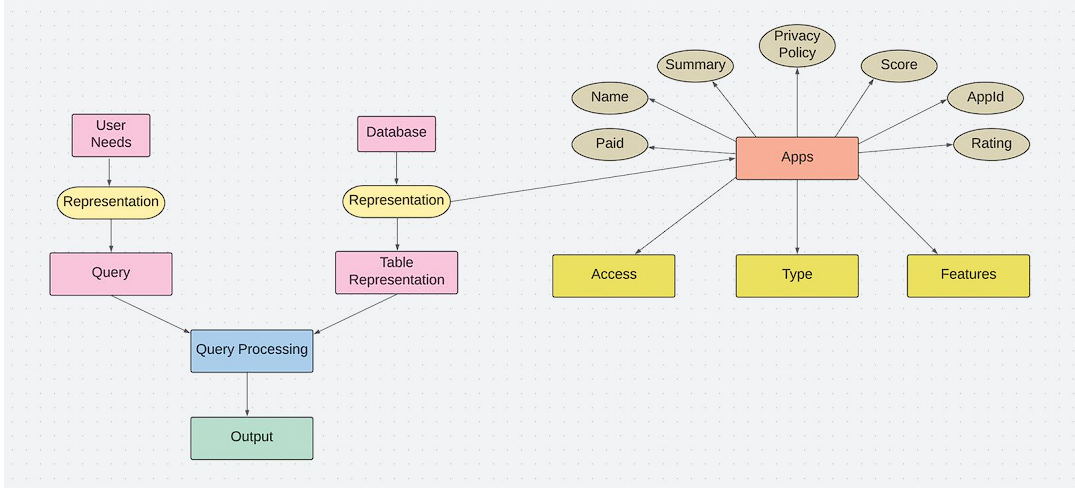
Furthermore, there are app recommendation systems like AppCrawlr, which suggest applications across different platforms based on user-desired features. If feasible, this concept could be extended to develop a recommendation system integrated with a summarizer, suggesting apps with similar features and comparatively more secure privacy policies. While these studies showcase the potential of leveraging privacy policies to identify unsafe apps and improve privacy practices, challenges persist regarding the accuracy of automated tools and the transparency and understandability of privacy policies.

## **Baseline Results** Data Collection

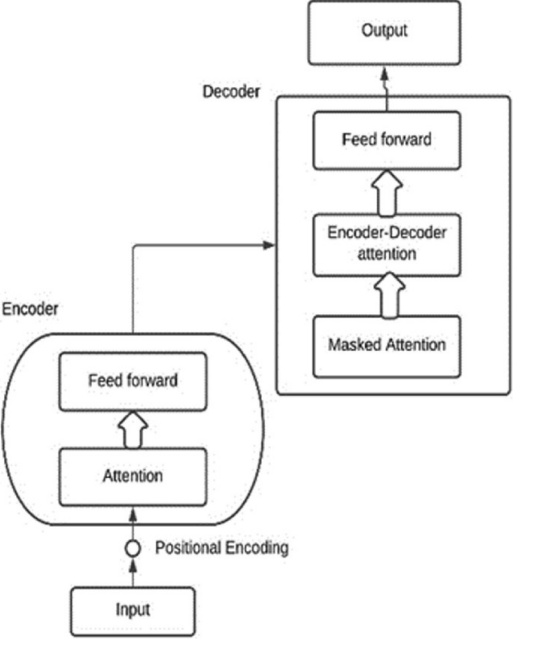
Data has been gathered from diverse origins. For the App-police application, the text for the Privacy Policy has been extracted from the official Privacy Policy document associated with the application on the Google Play Store. These privacy policy documents are also accessible online for numerous applications. Additionally, for our database, basic information regarding the permissions requested by the application upon installation is verified from the mobile device's settings. This data is stored in our backend database and utilized for subsequent analysis.

## Database





ER Diagram for the backend database  
  
Summarizations

In our database, we store information about the various permissions requested by applications, their Play Store ratings, and the features they offer. This data serves as the basis for our final application recommendations. Initially, we preprocess the data using techniques such as tokenization, lemmatization, and removal of punctuation marks. For text summarization, we employ the pre-trained "Transformer" model T5-BASE, developed by Google. T5, which stands for Text-to-Text Transfer Transformer, is a transformer-based language model trained on a vast corpus of text data and fine-tuned for various natural language processing (NLP) tasks, including question-answering, summarization, and language generation. The architecture of T5 is based on the transformer model, a neural network capable of processing sequential data like text. It comprises an encoder-decoder structure where the encoder converts the input sequence into a fixed-length vector representation, while the decoder utilizes the encoder output and a task-specific prompt to generate the final output sequence.  
  


## Score Calculations

We utilize weighted sums to assess the safety of the current app in our baseline evaluation. This involves categorizing the application into three groups: "secure," "unsafe," and "moderate." Subsequently, a set of keywords relevant to user privacy, such as "data collection," "third-party sharing," "advertising," etc., is identified. The privacy policy summary of the app is then extracted. To elaborate, we compute the frequency of each keyword and assign weights to them according to their significance in safeguarding user privacy. The overall app score is determined by the weighted sum of occurrences of these keywords.

## References

1. Vallina-Rodriguez, N., Balebako, R., Moreno, M., Grinberg, Y., Almuhimedi, H., Sundaresan, N., Felt, A., Aad, I., & Sadeh, N. (2015). Privacygrade: Measuring the privacy behaviors of smartphone apps. In Privacy Enhancing Technologies Symposium.

2. Cherivirala, S., Zhang, Y., Iyengar, A., & Cranor, L. (2017). Mining privacy policies for better mobile app design. IEEE Transactions on Software Engineering, 43(9), 834–848.

3. Nadkarni, A., Yadav, P., & Prakash, A. (2018). Privacy policy analysis of android VPN apps. In IEEE European Symposium on Security and Privacy Workshops.

4. Vallina-Rodriguez, N., Reaves, B., Shah, A., Sundaresan, N., Kreibich, C., Felt, A., & Paxson, V. (2017). Privacyscore: Analyzing the privacy behaviors of smartphone apps at scale. In IEEE Symposium on Security and Privacy.