A Step to Achieve Personalized Human Centric Privacy Policy Summary

Ivan Simon[1], Sherif Haggag[1] and Hussein Haggag[2]
[1]*Faculty of Science Engineering and Technology, The University of Adelaide, Adelaide, South Australia, Australia*
[2]*Department of Computing Science, Umeå University, Sweden*

Keywords: Privacy Policy, Text Summarization, Human-Centric, Machine Learning.

Abstract: Online users continuously come across privacy policies for the service they use. Due to the complexity and verbosity of policies, majority of the users skip the tedious task of reading the policy and accept it. Without reading and evaluating the document users risk giving up all kinds of rights to their personal data and for the most part, are unaware of the data sharing and handling process. Efforts have been made to address the complex and lengthy structure of privacy policies by creating a standardized machine-readable format of privacy policies for the web browsers to process it automatically, a repository of crowdsourced summarized versions of some privacy policies, or by using natural language processing to summarize the policies. PirvacyInterpreter is one unique tool that acknowledges human-centric factors while summarising the policy. Thus, it generates a personalised summary of the privacy policy for the user providing relevant information to appease their privacy concerns. This paper presents the conceptualization of PrivacyInterpreter and implements a proof-of-concept model using configured RoBERTa(base) model to classify a privacy policy and produce a summary based on privacy aspects that reflect users' privacy concerns.

1 INTRODUCTION

A privacy policy is a document that defines the practices of a company's data collection, use, and sharing practices. It legally binds a user who agrees to it. In the Internet-driven world, it is necessary to be wary of one's data as it could be used to achieve economic, political, and even nefarious gains. Privacy policies are notoriously unclear, complex, and wordy (Sigmund, 2021) (Steinfeld, 2016) (McDonald and Cranor, 2008). A recent data breach at OPTUS exposed sensitive data of 10 million customers, including past and present users. The customers were unaware that their data would be retained for a duration of 6 years. Such information is usually detailed in a privacy policy but with a fast-paced life, such document is overlooked by the user. According to the Australian Community Attitudes to Privacy Survey 2020, only 20% of Australians read and understand the privacy policies online (Lonergan Research, 2020).

The length and time to read the policies as well as the presence of obscure words have increased in the last 10 years, especially with the changes made by the Regulators like GDPR and CCPA (Amos et al., 2021). Tools have been developed to summarize privacy policies using machine learning (ML) algorithms. Privee (Zimmeck and Bellovin, 2014), PrivacyGuide (Tesfay et al., 2018), and PrivacyCheck (Zaeem et al., 2018) are some examples of the available summarization tools. They all have unique features. Privee calculates an overall grade to a policy helping users to understand quickly (Zimmeck and Bellovin, 2014), while PrivacyGuide classifies the privacy policies based on GDPR (Tesfay et al., 2018). Polisis is an automated framework for privacy policy analysis using neural network classifiers (Harkous et al., 2018).

It is found that the current summarization tools have limited performance (Bracamonte et al., 2019), lack personalization (Tesfay et al., 2018), and fail to consider the diversity in privacy concerns among users (Bergström, 2015a) (Škrinjarić et al., 2017a). These drawbacks motivate the development of a new form of policy summarization tool that acknowledges human-centric factors. The main contributions of the paper are:

1. Study of the diversity in online users privacy concerns.
2. Analysis of the current text classification models used to summarize privacy policy of websites.
3. To design a personalized privacy policy summarization tool and to develop its proof-of-concept model.

Simon, I., Haggag, S. and Haggag, H.
A Step to Achieve Personalized Human Centric Privacy Policy Summary.
DOI: 10.5220/0011842600003464
In *Proceedings of the 18th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2023)*, pages 381-395
ISBN: 978-989-758-647-7; ISSN: 2184-4895

The report is structured as follows: Section 3 contains a review of the literature exploring the diverse privacy concerns and analysis of the current text classification strategies used in summarization tools. Section 4 details the implementation of the proof-of-concept model. In Section 5, the results are presented and output analysis is given. Lastly, Section 6 concludes the paper, and future work is proposed in Section 7.

2 AIM

The current paper aims to seed the development of a human centric privacy policy summarization tool, i.e., PrivacyInterpreter. The output of this tool would address the concerns of the user and help them make an informed decision about their data.

3 LITERATURE REVIEW

3.1 Diversity in Online Privacy Concerns

People are diverse and have different aspirations and expectations when choosing a digital service. To predict user's privacy concerns researchers have looked into various factors like demography (Bergström, 2015b) (Lee et al., 2019) (Yao et al., 2007), personality traits (Taylor et al., 2015) (Wortman et al., 2012) (Škrinjarić et al., 2017b) (Haggag et al., 2022), the influence of the environment (Lee et al., 2016) and culture. Therefore, this study investigated the role of demographic characters and personality traits in users' privacy concerns.

Age: Age positively influenced online privacy concerns, people became more concerned about their privacy as age increased (Zhang et al., 2020). This may be attributed to increased awareness of Information Security as people grew old (McCormac et al., 2017). However, according to a recent study on the awareness of Single Sign-On (SSO) Security, Pratamal et. al. (Pratama et al., 2022) found that older people had less SSO security awareness as compared to young users. More precise statistics were given by Lee et. al. (Lee et al., 2016), that is, information privacy concerns were seen to peak in the age group of 20-30-year-olds and after the age of 40 the level of concern decreased. Privacy concerns about the use of social media decreased with age while concerns regarding the use of online debit cards for online transactions increased with age (Yao et al., 2007). Age was also closely related to the level of income and internet experience. In the age of the internet wherein the young generation are exposed to the internet very early, were found to have more internet experience than the older generation (Zukowski and Brown, 2007). Being more active online, they were highly likely to be victims of cybercrime (Näsi et al., 2015).

Mohammad and Schreuders (Schreuders et al., 2019) researched cybercrime victimization using police data and found that unwanted contact and online harassment targeted the younger generation. The possibility of being a target of such crime decreased with increasing age (30% for those aged 25-34 and 17% for those aged 35-44). The study showed that the highest targeted age group was 24-34 years. It was also seen that the victims of sexual and indecent image reception mostly fell in the age group of 5-14 years. According to Karagiannopoulos et. al. (Karagiannopoulos et al., 2021), people aged 60 years and above feared sharing personal information such as banking details and email addresses. Some of them were not well-versed with good cyber hygiene practices too. Young people who may be well educated about cyber hygiene were still not safe because they engaged in various online routines that increased their probability of encountering cyber fraud (this was true for educated people as well) (Whitty, 2019). For cyber scams, older adults were more prone to investment-related scams while young people were more prone to consumer-related scams (Whitty, 2020). Many online activities invited computer viruses to infect the system. The probability of obtaining a computer virus decreased with an increase in age (Ngo and Paternoster, 2011). This was true for cybercrime regarding online defamation too.

Gender: Research examining the role of gender in cyber security awareness suggested that females were more vulnerable than males to cyber threats. This suggestion was made either due to the low-security efficacy found in females (Anwar et al., 2017) or due to the underrepresentation of females in IT-related majors or work (Pratama and Firmansyah, 2021). Very small or no significant statistical difference was reported in the SSO security awareness (Pratama et al., 2022), privacy concerns (Näsi et al., 2015) and information security scores (McCormac et al., 2017) of males and females. Interestingly, it was observed that women have high information privacy concerns than men until 30 years of age and men after the age of 40 years showed more concern for their privacy than women (Lee et al., 2016). Women were more susceptible to phishing emails as compared to men (McCormac et al., 2017) (Halevi

et al., 2013) but it was also found that both men and women were equally skillful in identifying suspicious links in an email (Goel et al., 2017). *Therefore, as a precautionary step, it is important to warn women more about phishing emails than men.*

Cybercrimes related to unwanted contact, harassment, and obscene image reception victimized more females while fraud and theft targeted more males (Schreuders et al., 2019). Defamation and threat were among the most frequent cybercrimes (Näsi et al., 2015) to which males were commonly targeted. When it came to using credit/ debit cards online, women were more concerned than men about their online use (Bergström, 2015a). Men had a higher tendency to fall for investment scams than women and consumer scams victimized women more than men (Whitty, 2020). Whitty's (Whitty, 2020) research revealed that men and educated people were just as likely to be preyed on by cyber scams as women and less educated people. *Therefore, it is necessary to educate and caution men and women equally about cyber scams.* Using the ordered probit model check, Skrinjaric et. al. (Škrinjarić et al., 2017a) found that students were less concerned about their online privacy as compared to self-employed people. As people became more educated, their level of privacy concerns decreased. Education level and income positively influenced information privacy concerns (Lee et al., 2016).

3.1.1 Personality Traits and Privacy

Two of the Big Five personality traits were found to be closely related to online privacy concerns, i.e., "extraversion" and "neuroticism" (Škrinjarić et al., 2017a). The extrovert character was inversely related to online privacy concerns. A more extravert person was found to be less concerned. Openness was another driving trait that led people to publish information on social media sites and was correlated to less stringent privacy settings. Such activities suggested that the user may be less aware of the risks related to information leaks or the user may be blinded by the advantages of sharing the information on social media (Halevi et al., 2013). *Therefore, openness can be related to the disclosure of extra information*. Agreeableness, consciousness and neuroticism personality traits related to users who gave importance to privacy risks (Tang et al., 2020). A person who was moodier or had a high tendency of being vexed or was driven more by negative experiences was found to be more concerned about his/her online privacy (Tang et al., 2020). A high sense of SSO was found in people who were emotionally stable and conscious (Pratama et al., 2022). A high sense of SSO was found in people who were emotionally stable and conscious (Pratama et al., 2022). More agreeable people were more likely to have less awareness about security features like SSO but if they were to be more concerned about their privacy, then their behaviour was opposite (Pratama et al., 2022). People who had high confidence in their ability to cope-up with any situation had fewer online privacy concerns (Yao et al., 2007). More conscious and less risk-taking individuals had high information security awareness. Bergstrom et. al. (Bergström, 2015a), showed that trust in other people was an important determinant in understanding people's worries regarding unethical use of personal information while using emails.

Concerning cybercrime, the impulsive nature (Neupane et al., 2016) and emotional instability led to phishing victims and romance scams (Whitty, 2020). If self-control was described as the ability of an individual to resist in the face of temptation, then low self-control increased the chances of experiencing online harassment (Ngo and Paternoster, 2011). In another study it was seen that vulnerability to phishing attacks was inversely related to extraversion, and positively correlated to neuroticism and openness, especially for females (Halevi et al., 2013). High urgency and sensation-seeking behaviour (subcategories of impulsivity behaviour), correlated to engagement in routine activities that increased the chances of victimization by cyber frauds (Whitty, 2019). Counter-intuitively, Whitty (Whitty, 2019) also found that people who had perseverance, were educated and believed that events happened because of their actions and not by fate, were involved in online routines that made them more vulnerable to cyber fraud. Beyond Big-5 personality traits, variables like "privacy awareness and computer anxiety" positively drove online privacy concerns (Škrinjarić et al., 2017a). Privacy concerns regarding the collection of data was driven by compound traits "need for cognition" and "risk orientation" (Taylor et al., 2015).

3.1.2 Culture and Privacy Concerns

The cultural dimensions at the country level can be expressed broadly by using individualism versus collectivism. At a macroscopic level, collectivist culture values togetherness, group or community while individualistic culture gives importance to uniqueness and independence. It was found that collectivistic individuals gave more importance to group privacy and were more particular about the audience who could view their information than individualistic people (Li, 2022). In individualistic cultures, users were found to have a greater likelihood of sharing images and

videos on the public platform while collectivistic people were more comfortable sharing in only specific already involved groups like a workspace (Li, 2022).

Individualism was found to have more concern about individual privacy concerns and aspired for control strategies to protect personal privacy (Li, 2022). The authors showed that collectivistic individuals prioritise privacy risks in the use of Social Networking Sites (SNS) more than individualistic individuals (Trepte et al., 2017). Pertaining to the care free characteristic of individualistic culture, privacy implications of information shared by themselves or others could be suggested (Li, 2022). For example, when a user is sharing co-owned data, will all the co-owner be notified?// A study that surveyed participants from 38 countries, found that a lower level of Individualism, i.e., collectivism corresponded to a higher level of concern for errors in database storage (Bellman et al., 2004). But data collection by the government or users' employers is more supported by collectivistic cultures than individualistic cultures. The latter tend to put more trust in entities they already have a link with or have paid for (Li, 2022). Collectivistic users are more comfortable than individualistic individuals with data collection practices for automatic decision-making or customization (Li, 2022). While information on the involvement of a third-party in data collection could retract the consent of a user from collectivistic culture, the same information could make the individualistic user feel more comfortable about data collection (Li, 2022).

3.1.3 Attitudes of Australians Towards Data Privacy

The current research mains to develop a human-centric tool specifically for Australia, therefore it is necessary to understand what online privacy means to Australians. The Australian Community Attitudes to Privacy Survey 2020 (Lonergan Research, 2020) helped to understand the current dynamics of Australia's privacy concerns. Some of the key findings used in the research are detailed below.

Australians aged 18-24 years did not think much about data privacy (54%) when selecting an online service while those over the age of 65 years considered data privacy as one of the top 3 priorities (80%). Over 49% of the Australians aged 50 years and over were faced with unwanted communication for marketing purposes. The trend reduced by a small margin for individuals in the age group 35-49 years (42%) and reduced further to 35% for those aged 18-34 years. People aged 18-34 years actively reported about unnecessary data collection for a service. From a gender point of view, males (35%) reported more about unnecessary hoarding of data as compared to females (27%).

Identity theft, fraud and data breaches were considered the biggest risk by all the age groups with a slight reduction in percentage across younger groups. People even considered "sending information overseas" as a major risk. This risk prevailed in all age groups with 49% for 50 years and over, 35% for 35-49 years old and 34% for those aged 18-34 years. Digital practices like data sharing, location tracking, profiling, and use of AI also caused a high level of discomfort to the people. With location tracking, keeping databases of online activities and targeted advertisements in particular, older people (65 years and above) were more uncomfortable (75%) when compared to the younger generation (18-34 years old; 20%). But privacy concerns regarding social media decreased with an increase in age (overall percentage hovers around 14%). 66% of individuals were found to be more reluctant to provide biometric information compared to health information and location data. Surprisingly, Australians were comparatively comfortable providing other biometric information like voice prints (30%), facial images (35%) and fingerprints (43%) through smart devices and for government purposes. This comfort was only limited to a few organizations like the government and banking sector. An equal proportion of comfort and discomfort was seen for digital practices by the government like the use of facial recognition, video surveillance, identification of suspects, and use of biometrics.

Knowledge about privacy varied across age groups. 29% of the young generation (18-34 years of age) and 53% of early adopters were confident about their knowledge. Slightly fewer older Australians, i.e., aged 50 - 64+ years, rated their knowledge to be excellent or very good. Among the people who claimed to have excellent knowledge of data protection, 79% of them faced problems regarding the handling of PI. When it came to access to personal data, younger Australians were more dissatisfied compared to those aged 18-34 years and 35-49 years. A greater number of females and older Australians were uncomfortable sharing their location data when compared to males and those aged 18-49 years.

Australians considered the social media industry the least trustworthy, followed by search engines and apps. They had comparatively more trust in health care services, federal government departments, financial institutes and telecommunication providers. Compared to the younger generation, fewer people aged 35-49 years and 50+ years used privacy protection like the use of VPNs or privacy-preserving search engines. Females were more likely than males to ad-

just privacy settings on social networking websites and to change location sharing settings. But a greater number of males used VPNs, ad blockers and incognito mode than females.

The use of AI without the knowledge of the user was among the major privacy risk, especially for those aged 50 years and above (28%). This concern was slightly less for those aged 35-49 years (23%) and even less for 18-34-year-olds (20%). Younger Australians (31%) trusted the decision made by AI. This trust gradually decreased with the increase of age. An almost similar trend was seen in the use of AI for different purposes. People were more uncomfortable with the private sector than the public sector using AI. *Attitude towards children's privacy:* Australian parents were very concerned about their children's online privacy and expected that the online service or websites would take certain steps to protect the privacy of their children like: verification of age before collecting data, default privacy setting to be set to high-privacy mode and location tracking to be switched off for children, minimum collection of data.

Parents also struggled to find out ways to protect the personal information of their children whilst using the service. They found the task of comparing different privacy policies to select the best service or website difficult.

3.1.4 Cyber Threats and Privacy Policy

Common privacy-related cyber threats to Australia are: identify theft, cyber fraud/scams (consumer or investment related), online harassment (unwanted contact, illicit message or image reception), defamation and phishing emails. One of the causes of the aforementioned threats is the leak of data especially personal information (PI)(Lonergan Research, 2020). In any online activity a certain amount of user data is saved by the hosting organization. When PI is shared with a service provider it is always at a risk of being exposed to the public (Goel et al., 2017). If the data is not securely stored in the database and this data is breached, users' data is exposed to the world. Bad actors can use this publicly available data against the user to achieve their nefarious ends. Reyns (Reyns, 2013) through his research showed that routine activities like online shopping, banking, watching online shows or listening to music or news and online activities that involve sharing of PI with third parties (Reyns and Henson, 2016), are positively related to identity theft victimization. Thus, handling of users' data by organizations is an essential factor when dealing with identity theft (Milne et al., 2004), online fraud and scams.

The organizations must use appropriate security controls and impede violation of users' privacy. According to the 10 Generally Accepted Privacy Principles (GAPP), a privacy policy should contain the security measures adopted by the organization. This information would help users to weigh the risk involved in an activity (Balapour et al., 2020) and avoid dealings that could lead to cyber scams/ frauds.

Therefore, an online service provider must at the least use security controls like encryption (on both in-transit and stored data), network security, data breach response plan, access control (Culnan, 2019) and de-identification of data (Chang et al., 2018). If the data is outsourced for any purpose, it is necessary to implement the same level of security by the associated parties. Such measures could appease users' online concerns against activities like theft of data (or identity theft), online defamation and cyberbullying (data could be a user's possibly embarrassing facts) (Chang et al., 2018).

3.2 Text Classification Strategies

Traditional text classification models like the Naive Bayes, Random Forest, and Support Vector Machine, emphasize the feature extraction process (Solovyeva and Abdullah, 2022). Bag-of-Words or its variant term frequency-inverse document frequency lacks the ability to capture the semantics of words, creates a sparse matrix, and is computationally expensive (Verma et al., 2021). Though static word embeddings like Word2Vec and GloVe capture the embedding after analyzing the surrounding words, it generates only one embedding of the word irrespective of how the word is used in a sentence and ignores the possibility of words having more than one meaning (Verma et al., 2021). This leads to state-of-the-art models like BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) that use embedding like WordPiece (Wu et al., 2016). WordPiece and Byte-Pair Encoding learn the embeddings of subwords and capture the context in which a word is used. The embedding of subwords solves the problem of countering out-of-vocabulary (OOV) words by combining the representation of the subwords to form the OOV word. Some of the strategies used for summarizing and classifying privacy policies are detailed below:

1) PrivacyGuide (Tesfay et al., 2018)

It outputs the privacy categories, associated risk levels, along with the corresponding snippet of the policy.

In the preprocessing step, the StringToWordVector filter and Term frequency-inverse document frequency

(TF-IDF) was used. In addition to removing stop words, punctuation, tokenization, and stemming, the authors used a set of keywords to filter policy fragments to reduce the load on the classifiers. These keywords for privacy categories were discovered during the manual labelling of the training data set.

They used ML APIs from Waikato Environment for Knowledge Analysis (WEKA) and Java as the programming language. The prediction takes place in two parts, initially, the privacy aspect is predicted followed by the prediction of risk levels. On comparing the performance of PrivacyGuide using four different classifiers (namely, Random Forest, Decision Tree, Naïve Bayes, and Support Vector Machine.), the one with the Naïve Bayes engine performed the best. The authors also observed that the time for prediction decreased with every policy check but the time to collect and pre-process the policy was constant.

2) PrivacyCheck (Zaeem et al., 2018)

PrivacyCheck is a browser extension by nature that preprocesses the privacy policy captured from the input URL. The preprocessed text is sent to a data mining server to predict the risk level corresponding to 10 privacy factors (Zaeem et al., 2018). The client side was programmed using HTML and JavaScript (p. 53:5). A list of keywords was generated manually by analyzing 400 privacy policies (Zaeem et al., 2018). 11 data mining models were trained, ten for each privacy factor, and the eleventh to validate whether the URL corresponds to a privacy policy (Zaeem et al., 2018). The training data set was prepared by manually annotating 400 privacy policies. The Google Prediction API was used; therefore, the authors were unaware of the model used, as Google keeps the classification model confidential.

3) Privee (Zimmeck and Bellovin, 2014)

Privee is a browser extension written using JavaScript. The pre-processing stage is unique as it saves the position of the punctuations after removal and uses bigrams for better classification. Feature selection is conducted using regular expressions and used term frequency-inverse document frequency (TF-IDF).

The text prediction takes place in three steps. Initially, the text is subjected to regular expression of bigram. If a bigram matches, the text is classified using a rule-based classifier. Otherwise, in step two, a more general regular expression is used to extract the feature, and then stemming is done before input into the ML classifier. If no feature was extracted, it is concluded that the specific feature is not allowed or not present. For the design of the ML classifier, the WEKA library is used. After testing the various algorithms in WEKA, the naïve bayes algorithm in the multinomial version was selected. The multi-label classifier is divided into multiple instances with one classifier per category.

4) Polisis (Harkous et al., 2018)

Polisis annotates privacy policies into 10 high-level and 122 fine-grained classes using hierarchical neural network classifiers. The design of Polisis utilizes an application layer that acts as a front end for the reception and resolution of a user query. A data link layer scrapes the privacy policy and pre-processes it before feeding that queried privacy policy to the ML layer. The authors confirmed that when a webpage contains dynamic content, this dynamic content is already loaded into the webpage with the rest of the static content. Therefore, scraping a webpage after it is fully loaded gives the entire content of the webpage. Special steps were taken to prevent noisy annotation due to list aggregation. The acquired content is initially segmented according to HTML div-tags and p-tags, and then again subdivided using an unsupervised machine learning algorithm using domain-specific word embeddings.

The creation of domain-specific word embedding is considered the first stage of the machine learning layer. In the second stage, this word embedding is used to train a neural network. To train word embedding, fastText (a model that considers the composition and derivation of words) was used on a corpus of complied by crawling 130K privacy policy of application from Google Play Store. Polisis uses Convolution Neural Networks for multi-label classification at two levels. One level is to classify high-level privacy categories and the other is to classify values for each privacy attribute. The OPP-115 data set was used to train the classifiers. The categories and attributes correspond to the ones used in OPP-115.

5) BERT, RoBERTa, and PrivBERT

BERT (Bidirectional Encoder Representations for Transformer) is a multi-layer bidirectional transformer encoder. It is pretrained on unlabelled data using Masked LM (MLM) and Next Sentence Prediction (NSP) (Devlin et al., 2018). In MLM task the model must process the entire sequence before predicting the masked token, giving a better understanding of the context and flow of the language. The NSP helps the model to learn the relation between two sentences. The pre-training is done on a large dataset containing BooksCorpus and English Wikipedia was used (Devlin et al., 2018). Fine-tuning involves adding a layer on top of the core model. Task-specific input and output are given to the model, and the parameters are fine-tuned on the downstream task.

RoBERTa(Robustly optimized BERT) is a BERT model with an optimized pre-training method. It uses

dynamic masking, i.e., changing the masking in a sequence every time it is fed to the model (dynamic masking) (Liu et al., 2019). Byte-Pair encoding is used instead of WordPiece, the NSP task is removed, training is done over a large dataset for bigger batches (maximum steps= 500k with a batch size of 8K) and the input sequence length is longer than the original BERT (Liu et al., 2019).

PrivBERT is developed by using the pretrained RoBERTa(base) and retraining the model on privacy policy documents. But the vocabulary developed by the byte-level tokenization of the RoBERTa is retained, and out-of-vocabulary words were later fine tuned to incorporate the privacy jargon (Srinath et al., 2020). For fine-tuning the model, the OPP-115 annotated corpus was used. The performance evaluation of PrivBERT for the classification of data practices in a privacy policy showed improvement over Polisis and RoBERTa (Srinath et al., 2020).

3.3 Training Dataset

1) OPP-115 (Wilson et al., 2016).
The OPP-115 corpus contains 115 privacy policies with annotations at two levels. It contains 23K annotated data practices First, segments of paragraph size are annotated corresponding to 10 privacy categories (high level) followed by annotations of attributes: value pair (example: information type: email address) for each privacy category. In total, there are 10 high-level categories and 20 attributes with 138 different values.

2) PrivaSeer (Srinath et al., 2020).
It is a corpus of 1005380 privacy policies. The authors used Scrapy to crawl websites and Langid (python package) to filter the documents into 97 languages (p. 6831). Boilerplate another python package was used to remove the extra content of the web page like the header, footer, and navigation tabs. To classify English privacy policy from the scraped 3.2 million URLs, four machine learning models were trained, that is, three random forest models and fine-tuned a pretrained transformer-based language model (RoBERTa) (p. 6832). The size of the database is approximately 64GB when uncompressed.

3) CC-NEWS (Nagel, 2016).
CC-NEWS is a dataset from the Common Crawl. It crawls news websites around the world every day using StromCrawler (https://stormcrawler.net/). The data set contains news articles and is available on AWS S3. The raw crawled data are presented in WARC format (Web ARChive). The data set also stores metadata (WAT format) and text data (WET format) of the crawled sites. The size of the complete data set is in petabytes and contains data collected over 12 years.

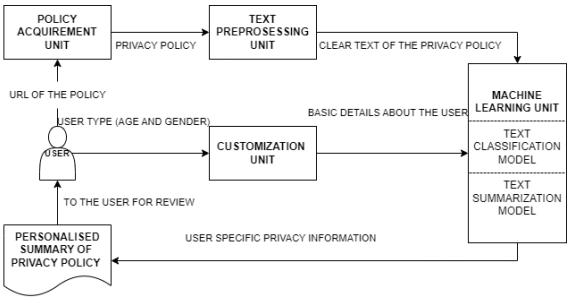


Figure 1: Tool Design.

4 METHODOLOGY

4.1 Privacy Categories

Based on the findings, detailed in the literature review, the following privacy categories have been created To add human-centric aspects to a privacy summarization tool:

Category 0: Australians

1. What Personal Information (PI) is being collected? How it is collected, stored and protected?
2. Whether or not PI is shared with overseas organizations? Which country does the sharing organization belong to?
3. Reason for collecting the PI.
4. Possibility of user profiling for targeted advertisement or business analysis purposes (it covers the data sharing practices).
5. Duration of holding the PI.
6. How will a breach of data be handled?
7. How to access given personal data?
8. How to complain about a privacy breach? Is there any practice that is exempted from The Privacy Act?
9. How to deal with an organization without giving away PI, i.e., in an anonymous manner?

Similarly, additional rights that Australians seek are:

1. Right to ask a business to delete their PI
2. Right to ask for compensation in court for a breach of privacy
3. Right to know about the use of PI in automated decision making

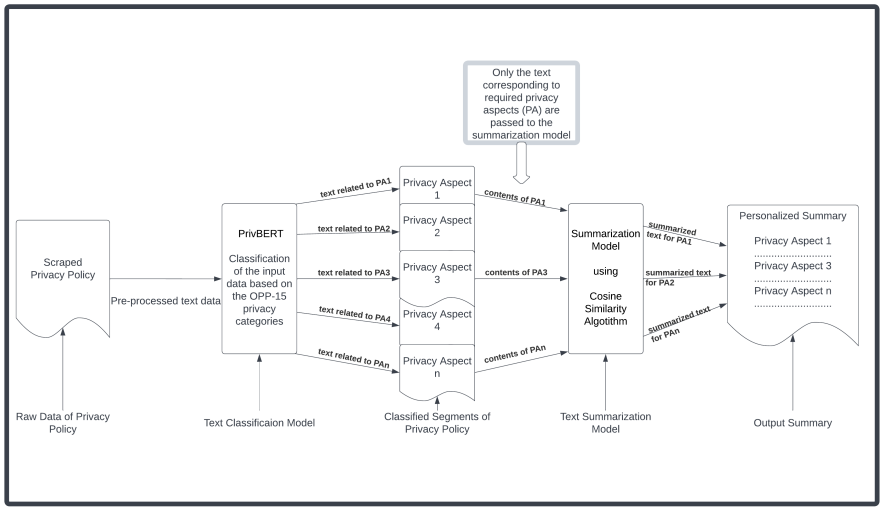


Figure 2: Policy Summarization Logic.

4. Right to object to certain data practices but still be able to use the services of the organization

Necessary warnings against:

1. identify theft and cyber fraud
2. storing user's information overseas
3. use of AI by the private sectors
4. government sector sharing information with private sector
5. insecure storage of user's data
6. the services related to social media and insurance company
7. third-party data collection if present

Category 1: Young

As privacy is not their major concern when selecting an online service or website, it is important to keep the summary succinct, highlighting their major concerns. Necessary details to be mentioned are:

1. irrelevant data collection
2. the possibility of targeted advertisements
3. whether and how the user can access their (supplied) personal information
4. specific cybercrimes to be notified are:
   Unwanted contact and online harassment, reception of Illicit images, consumer-related scams, possibility of downloading a virus.

Category 2: Old

Majority of the old people were found to have less exposure to the internet and have more concerns compared to the younger generation. The focus points based on their concerns found in the literature are

1. the possibility of unwanted communication for marketing purposes and targeted advertisement
2. unexpected information collection
3. location and online activity tracking
4. the collection of banking details
5. combining and sharing personal data among organizations.
6. disclosure of data supplied to another user
7. storing of biometric information in the device or organization's database
8. the use of AI in decision making along with what part of their PI will be used in the AI algorithm and whether a human is reviewing the decision made by the AI.
9. Specific cybercrime to be notified is:
   Investment-related scams

Category 3 and 4: Women and Men

Very faint distinct privacy concerns of men and women are found in the literature. This might be because a user's online privacy concerns are driven more by age and personality rather than their gender. Women are found to be vulnerable to similar privacy-related cyber threats as the young generation.
Necessary information for Women to be included are:

1. the collection of location details
2. the collection of banking details
3. the use of AI in the decision-making process
4. specific cybercrimes to be notified are:
   Unwanted contact and online harassment, obscene image reception, consumer-related scams and phishing emails.

Necessary information for Men to be included are:

1. irrelevant data collection
2. status of default privacy setting
3. specific cybercrimes to be notified are:
   Online fraud and theft, defamation and threat and investment-related scams

Category 5: Parents

Parameters that parents need to know regarding their children's privacy are:

1. location tracking by organizations
2. profiling (includes both collecting and inferring sensitive information) and selling data to third parties
3. targeting advertisements based on online profiling
4. unnecessary information about children being asked when a parent wants to use a service

Category 6 : Personality Traits

1. Extraversion: A person with high extraversion characteristics should be made aware of the consequences of sharing too many details online as they are less concerned about their privacy. They should also be educated about security features like SSO. A person with low extraversion should be cautioned about the possibility of phishing attacks.
2. Neuroticism: Warning about cyber scams and phishing attacks should be given to people with high neuroticism.
3. Openness: People who score high on this feature should be made aware of the risk of information leak as they, like highly extroverted persons, freely share information online. They should be informed about the different ways to harden privacy settings. As openness is directly related to phishing vulnerability, a person with high openness should all the more be warned about phishing attacks.
4. Impulsivity: Highly impulsive people should be cautioned about phishing attacks, cyber scams and cyber fraud.
5. Need for cognition: People who are highly cognitive are more sensitive towards data collection practices, hence they should be warned about the same.
6. Risk orientation: Like the need for cognition, a low-risk taker should also be warned about data collection practices.

Category 7 : Miscellaneous

1. Native/Non-Native: People who are non-native should be cautioned about online fraud and scams more prominently than native people.
2. Education level: Students should be educated and informed about the security features like Single-Sign-On and Two-Factor Authentication.

The high-level design of the policy summarization tool is shown in figure 1. To develop the proof of concept a mock customization unit is used in which four user types are present, each with 5 privacy aspects. The modified tool consists of three main components:

1) Policy Acquirement Unit (PAU):

The PAU scrapes the content of the privacy policy using Selenium and the Beautiful Soup library from the URL given by the user. After manually analyzing several online privacy policies, it was found that the main content is present under 'p' tags (including the 'li' and 'a' tags within the p tags). Therefore, the data within the 'p' tags and consequently 'li' and 'a' tags were scraped. To deal with dynamic pages, the unit waits till all the JavaScript code is loaded and then scrapes the content. At the same time, another approach was used, in which all embedded URLs were collected first from the page and then crawled one by one.

2) Text Preprocessing Unit (TPU):

It takes the scraped data and converts it into a dictionary of lists. The data are checked for duplicate sentences and then tokenized into separate words. The list of tokenized words is checked and sentences with less than 3 words are removed. This procedure removes redundant sentences like the language option, "Got to top " and "Contact Us".

3)Machine Learning Unit (MLU) (figure 2):

It is the core unit that generates the summary of the privacy policy. The summarization takes place in two stages.

- The scraped privacy policy is classified with the help of a text classification model. The output of the classification model contains segments of privacy policy corresponding to different privacy aspects.
- Only the segments of the privacy aspects that suit the user are selected and fed to another text summarization model to generate readable snippets of the information regarding the privacy aspects. The resultant snippets are combined to form the personalized summary of the privacy policy for the user.

Text Classification Model: For text classification, the paper uses the PrivBERT language model. It is based on the RoBERTa base model (Liu et al., 2019)

that is pre-trained with a dynamic Mask Language Modelling (MLM) task on a large data set of approximately 160 GB (Srinath et al., 2020). Like the RoBERTa base, PrivBERT also accepts a maximum of 512 tokens per input. To tokenize the training data, Byte-Pair Encoding is used with a vocabulary of size 50k. The HuggingFace transformer (Wolf et al., 2019) library provides the tokenizer and the basic PrivBERT model.

To train this model, the OPP-115 corpus (Wilson et al., 2016) was used. The "consolidated" subdirectory in the OPP-115 corpus contains CSV files of the annotated privacy policies with redundancies removed (since more than one annotator worked on each policy). The corpus contains paragraph-sized segments classified into 11 privacy categories. The dataset is structured in Pandas dataframe format with columns representing the text and privacy category. The data set was split into a 3:1: 1 ratio for training, testing, and validation.

Text Summarization Model: An extractive summarization technique is used employing Term Frequency- Inverse Document Frequency (TF-IDF) vectorization and cosine similarity. The cosine similarity gives the similarity between two sentences in vector form by calculating the angle between the two vectors. At first, the TF-IDF weights for each word in a sentence are calculated after excluding the stop words. The TF-IDF weights for each sentence are vectorized and are used to calculate the cosine similarity between pairs of sentences.As cosine is calculated, two sentences with a smaller angle will have a higher similarity score. The summary of the privacy aspect is created by selecting the top five sentences with the highest similarity score. Similarly, the summary of all the privacy aspects is generated and then compiled together to form the output summary.

5 RESULTS AND ANALYSIS

5.1 Results

At first, 600 samples from the training and validation data set were taken to train and validate the model. The model was trained with a batch size of 8 and a learning rate of 2e-5 for 10 epochs. The loss and accuracy were calculated over the validation set after the completion of each epoch. An accuracy of 0.7117 was seen for all 10 epochs with an average loss of 0.9121 over the validation set. The second time 2638 samples were used to train and 647 samples to validate the model. The model was trained for 10 epochs but with a batch size of 12 and a learning rate of 2.5e-5 (inspired by PrivBERT (Srinath et al., 2020)). An improved performance was seen after the completion of 10 epochs. The model attained an accuracy of 0.7814 and a loss of 0.6652 over the validation set (figure 4).

Sample run: The URL of the Stack Overflow website was provided and the user type was selected as "young female". The classification model took 1.50 minutes, and the time taken to extract similar text form the classified segments to form the summary was negligible. The scraped data consisted of 7199 words, and the output summary consisted of 908 words. The pre-selected privacy aspects for "young female" were "Policy Change", "International and Specific Audiences", "Practice Not Covered", and "Do Not Track". Sample run of more websites are summarised in figure3 The average time to classify a policy into 12 labels and to select 5 essential privacy aspects was found to be less than 2 minutes. Appendix section contains the model metrics after each training epoch and the privacy policy summary output.

5.2 Output Analysis

The absence of a privacy policy corpus with microlevel annotations to cover the wide range of privacy aspects used in the conceptualization of PrivacyInterpreter restrained the current implementation to classify the policy into 12 privacy aspects. These privacy aspects reflect the labels present in the annotated OPP-115 corpus, using which the classification model (PrivBERT) was trained. To demonstrate customization of the policy summary, a dummy customization function is created. The function takes a string input representing the age and gender of the user and outputs a list of IDs corresponding to the privacy aspects.

To calculate the execution time, an ipython-autotime module is used. It calculates the execution time for the code in each cell of the Google-Colab notebook. It was observed that the time taken to scrape a privacy policy and classify it into segments was the limiting factor to the overall time taken to generate the output summary. Other preprocesses and even the extractive summarization algorithm took less than a minute to execute. The time taken to scrape a webpage varied according to the size and design of the webpage. The output of the scraped data is essential, as it acts as the building block of the output policy summary. The website scraping code obtains the HTML document efficiently, but more refinement is required to obtain data between different tags (such as 'li' and 'a' tags) in a well-formatted way. Also,

when crawling the embedded links within a page, many links were found to be redundant. The presence of links such as 'back to top' or links presenting the same content in a different language increased the size of the input data. Thus, the crawling process is removed and the page is scraped only after all the JavaScript code is loaded to ensure that the complete content from the privacy policy page is captured.

Although state-of-the-art transformer models generate an abstractive summary, it is limited by the size of the input data and the output summary. For example, the Pegasus model (Zhang et al., 2020) gives optimal results but is restricted to an input token size of 512 and produces an output of 256 tokens. Therefore, more steps would be required to pre-process and split the classified segments obtained from the classification model into the acceptable input size of the Pegasus model. The use of cosine similarity for summarization overcomes this drawback. There is no restriction on the number of sentences. The input sentences are tokenized, vectorized, and the similarity score is calculated. The sentences with the highest similarity score are selected to form the output. Thus, the cosine similarity algorithm for text summarization also gives the flexibility to adjust the number of sentences in the output summary.

URL	User	Time $_{scrape}$	Time $_{classify}$	Words $_{initial}$	Words $_{final}$	Compression Ratio
https://stackoverflow.com/legal/privacy-policy	Young Female	2.56 seconds	1.50 minutes	7199	908	0.1261
https://policies.google.com/privacy?	Young Male	5.05 seconds	1.56 minutes	8734	1358	0.1555
https://legal.yahoo.com/au/en/yahoo/privacy/products/searchservices/index.html	Young Female	1.36 seconds	20.5 second	1229	60	0.0488 Note = Details of 3 categories out of 5 were not present in the policy
https://meta.wikimedia.org/wiki/Privacy_policy	Old Female	1.88 seconds	1.40 minutes	7722	948	0.1228
https://docs.github.com/en/site-policy/privacy-policies/github-privacy-statement	Old Female	2.09 seconds	1.54 minutes	6892	1129	0.1638
https://www.apple.com/legal/privacy/en-ww/	Old Male	4.16 seconds	1.25 minutes	7964	629	0.0790
https://www.linkedin.com/legal/privacy-policy	Young Female	2.45 seconds	1.56 minutes	12689	738	0.0582

Labels: URL = URL of the privacy policy
$Time_{scrape}$ = Time taken to scrape the privacy policy from the URL
$Time_{classify}$ = Time taken to classify the privacy policy
$Words_{initial}$ = Initial number of words in the privacy policy
$Words_{final}$ = Number of words in the output summary
Compression Ratio = Word count in summary / Word count in input privacy policy

Figure 3: Tool Performance Analysis.

5.3 Comparison with Existing Policy Summarization Tools

1. 1) PrivacyCheck (Zaeem et al., 2018)
   This tool uses a classification data mining model to inform the users about the level of risk involved in factors relating to the privacy and security of the user. It checks for 10 privacy factors and for each factor it asks some basic questions. Aspects covered by the privacy factors are based on a study of privacy protection guidelines provided by the Organization for Economic Co-operation and Development (OECD), Federal Trade Commission (FTC), and a survey interview of 16 employees and graduates (from the Center for Identity at UT-Austin) working in the field of security and privacy. It is logical to get advice from the experts and to refer the governing rules to design a tool but it is equally important to embody the diversity of the users in the design. For example, a website's data collection policy for children under 13 years might not interest a college student or a young user might not consider sharing an email address with a service provider as a risk. As mentioned earlier the length of the policy is one of the demotivating factors, it might be superfluous to add such information. Consider another example, Australians are very concerned about overseas data storage practices. If a privacy summary did not address such essential concerns, it is natural that an Australian user would deem the summary not useful and would not resort to using the privacy summarization tool again.

2. Privee (Zimmeck and Bellovin, 2014)
   It combines the use of a crowdsourced repository of summarized privacy policy along with Rule and ML classifiers. The limitations of this tool is two folds. Firstly, if the summary is gathered from the crowdsource repository, like "ToS;DR", it will be subjective to the competence and discretion of the person evaluating the policy for the repository. Secondly, the policy summary created using the Rule and ML classifiers is based on six binary categories (evaluated using yes or no). Such binary classification fails to give details like type of PI collected, duration of data retention or type of data protected, i.e., stored or in-transit data. Along with the summary, Privee also assigns an overall grade to the policy based on the outcome of the binary classifiers. The privacy categories just skim through the surface of the privacy policy and do not analyse the tricky bits. For example, the purpose of collecting the PI, where is the data stored, how would a user access his/her data, information about data control, and options to opt out of a particular service after initial sign-up.

3. PrivacyGuide (Tesfay et al., 2018)
   It uses ML and Natural Language Processing (NLP). The categories for the classification of the privacy policy are derived from the EU GDPR. Like PrivacyCheck, PrivacyGuard also uses a risk scale to inform the user about the level of risk involved. Based on the interpretation of the EU GDPR, the tools uses 11 privacy categories. These categories overlap with the privacy aspects used in this study. However, certain concerns specific to Australians are not covered by the privacy categories. For example, the possibility of targeted advertisement, the use of AI in decision making, the sectors involved in data sharing, information overseas storage of data practice and ways of dealing anonymously with the service provider. In all the three privacy summarization tools, the user is not warned about the innate cyber security attacks or threats a service might possess.

6 CONCLUSIONS

The paper presents the conceptualization of a personalized privacy policy summarization tool. To develop a proof-of-concept model PrivBERT, a RoBERTa-based language model pre-trained on a large privacy policy corpus, is further trained on a downstream task of text classification with the help of the OPP-115 (Wilson et al., 2016) privacy policy corpus. A dummy customization is used to represent the varied privacy aspects of online users. The classified segments of the privacy policy are condensed with the help of the cosine similarity algorithm to generate the output summary.

Despite the limitations encountered due to the scarcity of fine-level annotated privacy policy corpora and computational resources, the proof-of-concept model generates a summary in approximately 2 min with a compression ratio of 0.108 approximately (where compression ratio is defined as the ratio of the number of words in the output summary to the number of words in the input data). An online user could effortlessly read the generated summary which would appease their privacy concerns. The paper also suggests that incorporating human-centric aspects into design improves the productivity and utility of technology. The tool acts as a basic introductory version of a personalized summarization tool that could be reproduced, experimented with, and modified to develop better human-centric technologies.

7 FUTURE WORK

From the current vantage point, a corpus of privacy policies annotated at the micro-level is required, i.e., incorporating more diverse privacy aspects. Such a data set would help the summarization tool cover a wider variety of privacy aspects and produce a more personalized summary.

An efficient web crawler that is able to scrape data while maintaining the structure of the sentence or the format in which the sentence is presented on the website would improve the semantics of the output summary. Providing additional features, such as a presentation unit, to make the output more appealing or summarizing the 'Terms and Conditions' could also attract more users and enhance their awareness of online privacy.

Finally, the presented tool could be used as a starting point for the development of a personalized summarization tool incorporating more complex machine learning models like Generative Pre-trained Transformer models (version 3) and abstractive summarization using BART (Denoising Autoencoder for pretraining sequence-to-sequence) models.

REFERENCES

Amos, R., Acar, G., Lucherini, E., Kshirsagar, M., Narayanan, A., and Mayer, J. (2021). Privacy policies over time: Curation and analysis of a million-document dataset. In *Proceedings of the Web Conference 2021*, pages 2165–2176.

Anwar, M., He, W., Ash, I., Yuan, X., Li, L., and Xu, L. (2017). Gender difference and employees' cybersecurity behaviors. *Computers in Human Behavior*, 69:437–443.

Balapour, A., Nikkhah, H. R., and Sabherwal, R. (2020). Mobile application security: Role of perceived privacy as the predictor of security perceptions. *International Journal of Information Management*, 52:102063.

Bellman, S., Johnson, E. J., Kobrin, S. J., and Lohse, G. L. (2004). International differences in information privacy concerns: A global survey of consumers. *The Information Society*, 20(5):313–324.

Bergström, A. (2015a). Online privacy concerns: A broad approach to understanding the concerns of different groups for different uses. *Computers in Human Behavior*, 53:419–426.

Bergström, A. (2015b). Online privacy concerns: A broad approach to understanding the concerns of different groups for different uses. *Computers in Human Behavior*, 53:419–426.

Bracamonte, V., Hidano, S., Tesfay, W. B., and Kiyomoto, S. (2019). User study of the effectiveness of a privacy policy summarization tool. In *International Conference on Information Systems Security and Privacy*, pages 186–206. Springer.

Chang, Y., Wong, S. F., Libaque-Saenz, C. F., and Lee, H. (2018). The role of privacy policy on consumers' perceived privacy. *Government Information Quarterly*, 35(3):445–459.

Culnan, M. J. (2019). Policy to avoid a privacy disaster. *Journal of the Association for Information Systems*, 20(6):1.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Goel, S., Williams, K., and Dincelli, E. (2017). Got phished? internet security and human vulnerability. *Journal of the Association for Information Systems*, 18(1):2.

Haggag, O., Grundy, J., Abdelrazek, M., and Haggag, S. (2022). A large scale analysis of mhealth app user reviews. In *Empir Software Eng 27, 196 (2022)*.

Halevi, T., Lewis, J., and Memon, N. (2013). A pilot study of cyber security and privacy related behavior and personality traits. In *Proceedings of the 22nd international conference on world wide web*, pages 737–744.

Harkous, H., Fawaz, K., Lebret, R., Schaub, F., Shin, K. G., and Aberer, K. (2018). Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 531–548.

Karagiannopoulos, V., Kirby, A., Ms, S. O.-M., and Sugiura, L. (2021). Cybercrime awareness and victimisation in individuals over 60 years: A portsmouth case study. *Computer Law & Security Review*, 43:105615.

Lee, H., Wong, S. F., and Chang, Y. (2016). Confirming the effect of demographic characteristics on information privacy concerns.

Lee, H., Wong, S. F., Oh, J., and Chang, Y. (2019). Information privacy concerns and demographic characteristics: Data from a korean media panel survey. *Government Information Quarterly*, 36(2):294–303.

Li, Y. (2022). Cross-cultural privacy differences. In *Modern Socio-Technical Perspectives on Privacy*, pages 267–292. Springer, Cham.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lonergan Research (2020). Australian community attitudes to privacy survey 2020.

McCormac, A., Zwaans, T., Parsons, K., Calic, D., Butavicius, M., and Pattinson, M. (2017). Individual differences and information security awareness. *Computers in Human Behavior*, 69:151–156.

McDonald, A. M. and Cranor, L. F. (2008). The cost of reading privacy policies. *Isjlp*, 4:543.

Milne, G. R., Rohm, A. J., and Bahl, S. (2004). Consumers' protection of online privacy and identity. *Journal of Consumer Affairs*, 38(2):217–232.

Nagel, S. (2016). Cc-news.

Näsi, M., Oksanen, A., Keipi, T., and Räsänen, P. (2015). Cybercrime victimization among young people: a multi-nation study. *Journal of Scandinavian Studies in Criminology and Crime Prevention*, 16(2):203–210.

Neupane, A., Saxena, N., Maximo, J. O., and Kana, R. (2016). Neural markers of cybersecurity: An fmri study of phishing and malware warnings. *IEEE Transactions on information forensics and security*, 11(9):1970–1983.

Ngo, F. T. and Paternoster, R. (2011). Cybercrime victimization: An examination of individual and situational level factors. *International Journal of Cyber Criminology*, 5(1):773.

Pratama, A. R. and Firmansyah, F. M. (2021). Until you have something to lose! loss aversion and two-factor authentication adoption. *Applied Computing and Informatics*, (ahead-of-print).

Pratama, A. R., Firmansyah, F. M., and Rahma, F. (2022). Security awareness of single sign-on account in the academic community: the roles of demographics, privacy concerns, and big-five personality. *PeerJ Computer Science*, 8:e918.

Reyns, B. W. (2013). Online routines and identity theft victimization: Further expanding routine activity theory beyond direct-contact offenses. *Journal of Research in Crime and Delinquency*, 50(2):216–238.

Reyns, B. W. and Henson, B. (2016). The thief with a thousand faces and the victim with none: Identifying determinants for online identity theft victimization with routine activity theory. *International journal of offender therapy and comparative criminology*, 60(10):1119–1139.

Schreuders, C. et al. (2019). Understanding cybercrime victimisation: modelling the local area variations in routinely collected cybercrime police data using latent class analysis. *International Journal of Cyber Criminology*, 13(2):493–510.

Sigmund, T. (2021). Attention paid to privacy policy statements. *Information*, 12(4):144.

Škrinjarić, B., Budak, J., and Žokalj, M. (2017a). The effect of personality traits on online privacy concern. *Radni materijali EIZ-a*, (2):5–29.

Škrinjarić, B., Budak, J., and Žokalj, M. (2017b). The effect of personality traits on online privacy concern. *Radni materijali EIZ-a*, (2):5–29.

Solovyeva, E. B. and Abdullah, A. (2022). Comparison of different machine learning approaches to text classification. In *2022 Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*, pages 1427–1430. IEEE.

Srinath, M., Wilson, S., and Giles, C. L. (2020). Privacy at scale: Introducing the privaseer corpus of web privacy policies. *arXiv preprint arXiv:2004.11131*.

Steinfeld, N. (2016). "i agree to the terms and conditions":(how) do users read privacy policies online? an

eye-tracking experiment. *Computers in human behavior*, 55:992–1000.

Tang, J., Akram, U., and Shi, W. (2020). Why people need privacy? the role of privacy fatigue in app users' intention to disclose privacy: based on personality traits. *Journal of Enterprise Information Management*.

Taylor, J. F., Ferguson, J., and Ellen, P. S. (2015). From trait to state: Understanding privacy concerns. *Journal of Consumer Marketing*.

Tesfay, W. B., Hofmann, P., Nakamura, T., Kiyomoto, S., and Serna, J. (2018). Privacyguide: towards an implementation of the eu gdpr on internet privacy policy evaluation. In *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*, pages 15–21.

Trepte, S., Reinecke, L., Ellison, N. B., Quiring, O., Yao, M. Z., and Ziegele, M. (2017). A cross-cultural perspective on the privacy calculus. *Social Media+ Society*, 3(1):2056305116688035.

Verma, V. K., Pandey, M., Jain, T., and Tiwari, P. K. (2021). Dissecting word embeddings and language models in natural language processing. *Journal of Discrete Mathematical Sciences and Cryptography*, 24(5):1509–1515.

Whitty, M. T. (2019). Predicting susceptibility to cyber-fraud victimhood. *Journal of Financial Crime*.

Whitty, M. T. (2020). Is there a scam for everyone? psychologically profiling cyberscam victims. *European Journal on Criminal Policy and Research*, 26(3):399–409.

Wilson, S., Schaub, F., Dara, A. A., Liu, F., Cherivirala, S., Leon, P. G., Andersen, M. S., Zimmeck, S., Sathyendra, K. M., Russell, N. C., et al. (2016). The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1340.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Wortman, J., Lucas, R. E., and Donnellan, M. B. (2012). Stability and change in the big five personality domains: evidence from a longitudinal study of australians. *Psychology and aging*, 27(4):867.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yao, M. Z., Rice, R. E., and Wallis, K. (2007). Predicting user concerns about online privacy. *Journal of the American Society for Information Science and Technology*, 58(5):710–722.

Zaeem, R. N., German, R. L., and Barber, K. S. (2018). Privacycheck: Automatic summarization of privacy policies using data mining. *ACM Transactions on Internet Technology (TOIT)*, 18(4):1–18.

Zhang, J., Zhao, Y., Saleh, M., and Liu, P. (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

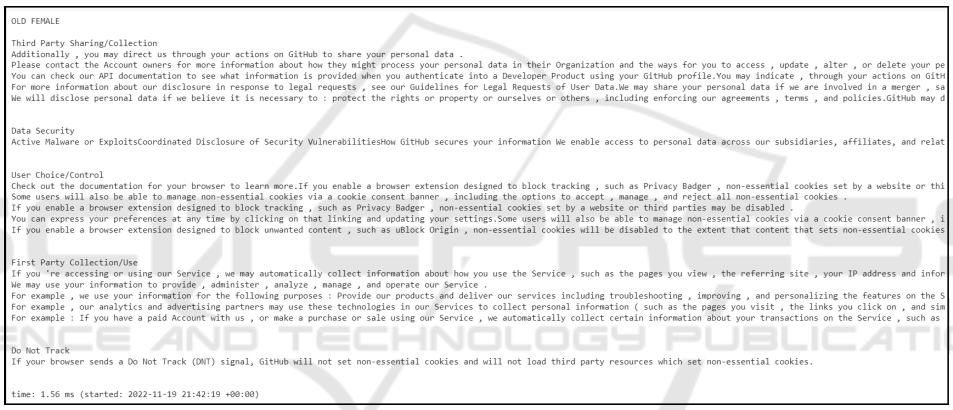
Zimmeck, S. and Bellovin, S. M. (2014). Privee: An architecture for automatically analyzing web privacy policies. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 1–16.

Zukowski, T. and Brown, I. (2007). Examining the influence of demographic factors on internet users' information privacy concerns. In *Proceedings of the 2007 annual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries*, pages 197–204.

APPENDIX

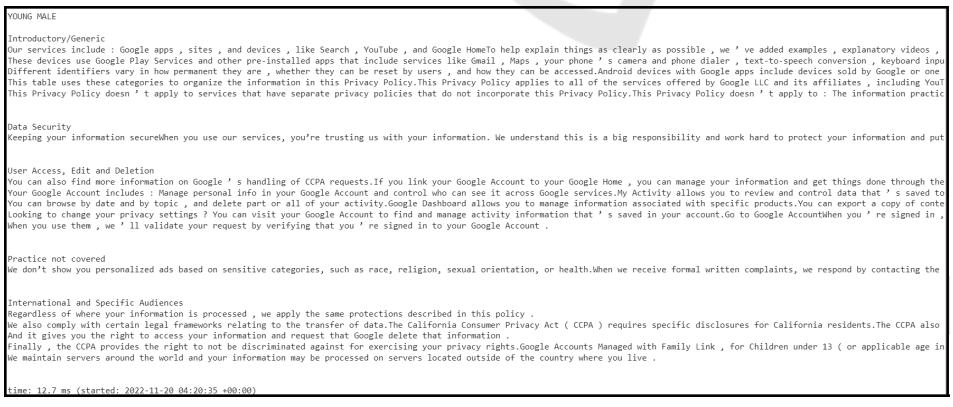
```
Epoch 1/10
219/219 [==============================] - 248s 1s/step - loss: 0.6387 - val_loss: 0.7024 - accuracy: 0.7673
Epoch 2/10
219/219 [==============================] - 248s 1s/step - loss: 0.5485 - val_loss: 0.6921 - accuracy: 0.7767
Epoch 3/10
219/219 [==============================] - 246s 1s/step - loss: 0.4855 - val_loss: 0.6791 - accuracy: 0.7657
Epoch 4/10
219/219 [==============================] - 252s 1s/step - loss: 0.4369 - val_loss: 0.6588 - accuracy: 0.7673
Epoch 5/10
219/219 [==============================] - 251s 1s/step - loss: 0.3935 - val_loss: 0.6724 - accuracy: 0.7783
Epoch 6/10
219/219 [==============================] - 251s 1s/step - loss: 0.3650 - val_loss: 0.6690 - accuracy: 0.7704
Epoch 7/10
219/219 [==============================] - 250s 1s/step - loss: 0.3405 - val_loss: 0.6534 - accuracy: 0.7799
Epoch 8/10
219/219 [==============================] - 252s 1s/step - loss: 0.3129 - val_loss: 0.6652 - accuracy: 0.7830
Epoch 9/10
219/219 [==============================] - 250s 1s/step - loss: 0.3023 - val_loss: 0.6652 - accuracy: 0.7783
Epoch 10/10
219/219 [==============================] - 248s 1s/step - loss: 0.3020 - val_loss: 0.6652 - accuracy: 0.7814
<keras.callbacks.History at 0x7f58e8a74d50>
```

Figure 4: Model Metrics when trained with 2683 samples, a batch size of 12 and learning rate of 2e-5 for 10 epochs.



```
OLD FEMALE

Third Party Sharing/Collection
Additionally , you may direct us through your actions on GitHub to share your personal data .
Please contact the Account owners for more information about how they might process your personal data in their Organization and the ways for you to access , update , alter , or delete your pe
You can check our API documentation to see what information is provided when you authenticate into a Developer Product using your GitHub profile.You may indicate , through your actions on GitH
For more information about our disclosure in response to legal requests , see our Guidelines for Legal Requests of User Data.We may share your personal data if we are involved in a merger , sa
We will disclose personal data if we believe it is necessary to : protect the rights or property or ourselves or others , including enforcing our agreements , terms , and policies.GitHub may d


Data Security
Active Malware or ExploitsCoordinated Disclosure of Security VulnerabilitiesHow GitHub secures your information We enable access to personal data across our subsidiaries, affiliates, and relat


User Choice/Control
Check out the documentation for your browser to learn more.If you enable a browser extension designed to block tracking , such as Privacy Badger , non-essential cookies set by a website or thi
Some users will also be able to manage non-essential cookies via a cookie consent banner , including the options to accept , manage , and reject all non-essential cookies .
If you enable a browser extension designed to block tracking , such as Privacy Badger , non-essential cookies set by a website or third parties may be disabled .
You can express your preferences at any time by clicking on that linking and updating your settings.Some users will also be able to manage non-essential cookies via a cookie consent banner , i
If you enable a browser extension designed to block unwanted content , such as uBlock Origin , non-essential cookies will be disabled to the extent that content that sets non-essential cookies


First Party Collection/Use
If you 're accessing or using our Service , we may automatically collect information about how you use the Service , such as the pages you view , the referring site , your IP address and infor
We may use your information to provide , administer , analyze , manage , and operate our Service .
For example , we use your information for the following purposes : Provide our products and deliver our services including troubleshooting , improving , and personalizing the features on the S
For example , our analytics and advertising partners may use these technologies in our Services to collect personal information ( such as the pages you visit , the links you click on , and sim
For example : If you have a paid Account with us , or make a purchase or sale using our Service , we automatically collect certain information about your transactions on the Service , such as


Do Not Track
If your browser sends a Do Not Track (DNT) signal, GitHub will not set non-essential cookies and will not load third party resources which set non-essential cookies.


time: 1.56 ms (started: 2022-11-19 21:42:19 +00:00)
```

Figure 5: Output of GitHub privacy policy for an Old Female.



```
YOUNG MALE

Introductory/Generic
Our services include : Google apps , sites , and devices , like Search , YouTube , and Google HomeTo help explain things as clearly as possible , we ' ve added examples , explanatory videos ,
These devices use Google Play Services and other pre-installed apps that include services like Gmail , Maps , your phone ' s camera and phone dialer , text-to-speech conversion , keyboard inpu
Different identifiers vary in how permanent they are , whether they can be reset by users , and how they can be accessed.Android devices with Google apps include devices sold by Google or one
This table uses these categories to organize the information in this Privacy Policy.This Privacy Policy applies to all of the services offered by Google LLC and its affiliates , including YouT
This Privacy Policy doesn ' t apply to services that have separate privacy policies that do not incorporate this Privacy Policy.This Privacy Policy doesn ' t apply to : The information practic


Data Security
Keeping your information secureWhen you use our services, you're trusting us with your information. We understand this is a big responsibility and work hard to protect your information and put


User Access, Edit and Deletion
You can also find more information on Google ' s handling of CCPA requests.If you link your Google Account to your Google Home , you can manage your information and get things done through the
Your Google Account includes : Manage personal info in your Google Account and control who can see it across Google services.My Activity allows you to review and control data that ' s saved to
You can browse by date and by topic , and delete part or all of your activity.Google Dashboard allows you to manage information associated with specific products.You can export a copy of conte
Looking to change your privacy settings ? You can visit your Google Account to find and manage activity information that ' s saved in your account.Go to Google AccountWhen you ' re signed in ,
When you use them , we ' ll validate your request by verifying that you ' re signed in to your Google Account .


Practice not covered
We don't show you personalized ads based on sensitive categories, such as race, religion, sexual orientation, or health.When we receive formal written complaints, we respond by contacting the


International and Specific Audiences
Regardless of where your information is processed , we apply the same protections described in this policy .
We also comply with certain legal frameworks relating to the transfer of data.The California Consumer Privacy Act ( CCPA ) requires specific disclosures for California residents.The CCPA also
And it gives you the right to access your information and request that Google delete that information .
Finally , the CCPA provides the right to not be discriminated against for exercising your privacy rights.Google Accounts Managed with Family Link , for Children under 13 ( or applicable age in
We maintain servers around the world and your information may be processed on servers located outside of the country where you live .


time: 12.7 ms (started: 2022-11-20 04:20:35 +00:00)
```

Figure 6: Output of Google privacy policy for a Young Male.