Investigating COPPA Notification Compliance

Are App Developers Ensuring Compliance for the Sake of Children's Privacy?

Mark Brom, Lydia Esbaum, Evan Lemker, Peter Mertka, Sebastian Rivera

ABSTRACT

In 1998, the United States Congress passed the Children's Online Privacy Protection Act (COPPA), the first bill that addressed the need to protect children's privacy. Due to children's young age and susceptibility to targeted advertising that stems from data collection, regulations must provide some rules and guidance regarding what can and cannot be collected from children. COPPA requires those who collect children's information to follow many different practices. Our study focuses on the notification of collection via an entity's privacy policy.

To comply with COPPA's regulations, three things must be displayed on a website regarding the data collection of children. First, there must be contact information, including the name, address, phone number, and email address of all parties collecting data or of one designated data monitoring officer. Second, there must be a clear description of what kinds of data are being collected from users who are thirteen years old or younger. Lastly, it is necessary that either a direct link to or a process describing how a parent can review or request deletion of their child's data must be available.

This study investigates compliance with the necessary notification requirements of COPPA by examining apps on the Google Play Store. The study collects the privacy policies of over five hundred apps found in the "Kids" section of the Google Play Store using a one-time web crawl. We analyze the content of the privacy policies to see which portions of notification are included and to what extent. The results of this privacy policy review provide a statistical understanding of how much of this bill is being followed twenty-five years after its initial passing for a medium that did not exist during the bill's initial passing.

KEYWORDS

COPPA, Privacy Policies, Google Play Store, Compliance

1 Introduction

On October 21st, 1998, the Children's Online Privacy Protection Act was signed into law, bringing necessary regulation and protection to reign in an ever-expanding and evolving online ecosystem. The Act went into effect on April 21st, 2000, allowing commercial websites and online services to ensure compliance with all the aspects of the bill [1]. Twenty-three years later, privacy is still a hotly debated issue and has become increasingly concerning as the number of avenues to collect data expands.

One avenue that had yet to exist at the time of the law's passage is that of mobile applications. With the rise of smartphones, tablets, and Internet of Things devices over the past ten years, it has become much easier for children below the age of thirteen to not only access the Internet but to use it without supervision. As a result, many mobile applications have been solely developed and targeted toward children. With apps of this kind, developers can collect vast swaths of information that can be leveraged for profit against the user via targeted marketing and user profiling. This practice, however, is the exact type of behavior that COPPA is designed to protect against. Thus, it is necessary to evaluate and ask just how many of these developers comply with the exact requirements of the law.

As new technologies evolve and multiply, which have outpaced our current legal constraints, one of the most prominent and popular mediums of internet access for children must follow through with the necessary compliance. By understanding where companies and developers are failing, it will be possible to provide

sound recommendations on how these policies can be better enforced for current and future technologies.

COPPA is a large and complex document that outlines numerous requirements and regulations that applicable companies must follow. However, it would be easier to identify some of the most explicit requirements of the Act. For this study, we focused on one portion of COPPA, specifically Section 312.4(d). This section was chosen because it lends itself to a simple auditing method. With the section listing three basic requirements that companies need to include in their privacy policy, we felt confident that we could devise a method to test for compliance with each subcomponent of the section. By focusing on Section 312.4(d), we can better identify compliance (or the lack thereof) while also considering the constraints imposed by this style of audit. A broader scope would lead to a less organized study considering the time requirements of this study. To better understand what we are investigating, it is vital to know what Section 312.4(d) entails. The exact text of the mandate is as follows:

312.4d (Official Statement): ***Notice on the Web site or online service.*** In addition to the direct notice to the parent, an operator must post a prominent and clearly labeled link to an online notice of its information practices concerning children on the home or landing page or screen of its Web site or online service, *and,* at each area of the Web site or online service where personal information is collected from children. The link must be close to the information requests in each area. An operator of a general audience Web site or online service with a separate children's area must post a link to a notice of its information practices regarding children on the home or landing page or screen of the children's area. To be complete, the online notice of the Web site or online service's information practices must state the following:

(1) The name, address, telephone number, and email address of all operators collecting or maintaining personal information from children through the Web site or online service. *Provided that:* The operators of a Web site or online service may list the name, address, phone number, and email address of one operator who will respond to all inquiries from parents concerning the operators' privacy policies and use of children's information, as long as the names of all the operators collecting or maintaining personal information from children through the Web site or online service are also listed in the notice;
(2) A description of what information the operator collects from children, including whether the Web site or online service enables a child to make personal information publicly available; how the operator uses such information; and the operator's disclosure practices for such information; and
(3) The parent can review or have deleted the child's personal information, refuse to permit further collection, or use of the child's information, and state the procedures for doing so [2].

In Section 312.4(d), websites or online services must include clear identifiable links tied to information notices regarding children on the site's home or other landing pages. Each site needs three components: 1) contact information of those collecting the data (names, address, email, and phone number); 2) a section stating what is being collected; and 3) a section devoted to the review and deletion of the child's data. Our web scrape of the Google Play Store looks for these policies and takes them from the web for further analysis. From there, our text segmentation scripts, and natural language processing model identify these three components and will determine whether an app complies with COPPA. If an app service needs one of the three components, we can prove that the app is not entirely COPPA-compliant. Throughout this process, we aim to answer the following research questions:

-**RQ1:** How many companies are posting the necessary data collector contact information so parents can contact the proper data authorities?

-**RQ2:** Which development companies are clearly defining what is being collected of children by their applications?

-**RQ3:** What portion of companies in our dataset are providing explicit and proper instruction for parents to review and delete their children's data?

Compliance must be analyzed because companies in the past have disregarded COPPA regulations, thus making it incredibly difficult for parents to assert necessary control and judgment on behalf of their child [3]. Reviewing and understanding what sort of information is being collected from children is vital to

ensuring that companies remain accountable to the regulations COPPA requires.

To answer these questions, we utilized a three-step process that can crawl various sections of the Google Play Store's "Kids" section. Firstly, we acquire the content of various children's applications' web pages from the Google Play Store website. This content contains all possible information about the app, including name, rating, developer, reviews, and images. From there, each page's results are scraped to isolate the company's directly posted privacy policy. From there, each policy is split into manageable one to three-sentence chunks so that our natural language model can appropriately classify it. Lastly, each chunk is parsed by a combination of pattern-matching expressions and our model to directly state the degree to which companies comply with each aspect of Section 312.4(d).

Overall, our results have shown that many companies need more compliance with the outlined rules, with only a tiny percentage of our companies in our data set fully compliant with all the aspects of COPPA that we are analyzing. However, while only some companies are fully compliant, there are still degrees of compliance in each of the three sections. This allows for further recommendation and emphasis on the sections with the least compliance, chiefly the lack of instruction on reviewing and deleting collected data. Companies are at least partially compliant by virtue of privacy policy design or conscious decisions.

2 Background

We begin our process by reviewing prior works that have either audited similar laws or investigated various aspects of COPPA.

2.1 Prior Work

COPPA is a set of requirements that online services must abide by that deal with managing data from children under 13 years of age. The purpose of COPPA is to protect children on the Internet by regulating what personal information online services can gather, use, and share about children. COPPA also gives parents control over how their children's personal information is gathered and shared by attaining verifiable parental consent. COPPA only applies to services that are either directed toward children under the age of 13, services that are directed to the public but know they have collected information from children under 13, or services that know they have collected information directly from users of another online service directed to children [4].

Several prior studies have analyzed COPPA and apps' compliance with COPPA. Our group has chosen to model our project loosely based on one of these studies, "'Won't Somebody Think of the Children' Examining COPPA Compliance at Scale." Researchers of this study created an automated evaluation framework for the privacy practices of Android apps. Specifically, the top 5,855 apps geared toward children that COPPA governs from Google's Play Store in the U.S. were used in the analysis. Unlike many approaches that aim to identify potential COPPA violations but fail to do so because they do not observe actual violations or do not scale, the framework used in this study allowed researchers to supervise apps' behaviors in real-time and at scale [4].

The study methodology included retrieving apps from a corpus of free, children-directed apps on the Google Play Store, running each app, and analyzing the information collected about each app's access to personal information and communication with third parties. During analysis, parsing and extracting certain pieces of information, like whether an app accessed Android-guarded resources, was an automated process, while obtaining other information, like checking for personal information in network transmissions, was manual. Like this approach, our group will automate parts of our analysis and manually analyze the data. We will also use children-directed apps from the Google Play Store in our project, and we will only examine 500 different applications. Our project complements this study by focusing on whether apps comply with a specific section of COPPA, Section 312.4(d) Notice on the website or online service, rather than analyzing if any section of COPPA is violated.

Another study that closely resembles the research we conducted is "Analyzing Privacy Policies Through Syntax-Driven Semantic Analysis of Information Types." This research paper focuses on creating a

program that automatically analyzes complex privacy policies and generates short summaries of what is being collected and shared in the policy. Our goal is similar as we want to create a program that can automatically analyze privacy policies to find the presence or lack of required components under COPPA regulation. While the formerly mentioned research project helps identify key concepts and categories relevant to users' privacy concerns, we aim to identify shortcomings in regulatory requirements relevant to companies covered by COPPA [5].

The researchers in this study designed a program using natural language processing techniques to perform syntax-driven semantic analysis of each part of the privacy policies. These techniques included heavily focusing on information types in the privacy policy, such as names, email addresses, phone numbers, locations, etc., to identify what pieces of information were being used and how they were being used. Similarly, we will have critical information we will be searching for within privacy policies, such as web links, email addresses, and phone numbers. However, rather than using natural language processing only to understand what the company does with this information, we will be using it to understand whether these links and email addresses can be used to access the online notices and data collection contact information that is required under COPPA for companies' privacy policies to contain.

One study that does not directly apply to what we are researching but is structured in such a way that it has served as a massive basis for the overall experimental design and construction of this project. Christo Wilson, a professor at Northeastern University, worked closely with a student, Maggie Van Nortwick, to develop a method for testing compliance with similar privacy law, the California Consumer Protection Act (CPPA). In their paper, "Setting the Bar Low: Are Websites Complying With the Minimum Requirements of the CCPA?" they set out to answer a similar question as in our work: just how much are companies complying with privacy laws [6]?

In their work, Wilson and Nortwick outline a complex method of scanning the web's top one million most popular websites to determine whether they were following one of the most basic requirements established by the CCPA, that being the need for a link stating "Do Not Sell My Private Information." This relatively simple requirement gave them a way to determine clear and defined compliance within the bounds of the law so that it could be better understood just how many companies were following specified guidelines [6]. This exact method and question formation inspired our group's motivation to investigate a similar question but through the lens of COPPA. Although the laws we are analyzing differ, the methodology of doing a web scrape before scanning the results for actual compliance with a specific mandate of the law served as a basis for our experiment.

Lastly, another research group at the University of Iowa: The Security, Privacy, and Anonymity Research Team, or SPARTA Lab, is investigating a different law using a similar method. This lab is headed by Dr. Rishab Nithyanand and is currently working on several projects related to online privacy and regulation. Within these projects, there is one aiming to analyze privacy policies just as we are for compliance with regulatory frameworks. Specifically, they use the same natural language processing guided approach we used in our research. However, instead of using it to determine compliance with COPPA, their methods are concerned with determining compliance with the CCPA. This is significant because we will be able to collaborate with members of the research team, such as Maaz Bin Musa, a Ph.D. candidate at the University of Iowa studying under the supervision of Dr. Nithyanand, to gain insights into the design and use of different natural language processing techniques to find the key results we are searching for.

2.2 Applicability

The first key question to ask before we began collecting privacy policies was to understand what entities COPPA applied to. Per the FTC's rules, COPPA encapsulates all websites and online services (such as mobile apps) directly targeted at children aged 13 or younger. Furthermore, anyone with knowledge of collecting, using, and disclosing children's personal information is included, even if the data is collected from a different site [1]. COPPA itself outlines who needs to comply with the specifications it sets forth.

Another critical distinction necessary for identifying to whom COPPA applies is that the country of origin of the website or app's controller does not exempt them from the law. If the website or service is targeted at and used by U.S. children 13 years old or younger, they must follow all requirements [1]. Thankfully, this meant that our web scrape and app selection could have considered whether the app itself was expected to comply with the law. The app is available in the "Kids" section of the U.S. version of the Google Play Store. That application is being targeted toward U.S. children. This is because the "Kids" section is explicitly for applications marketed to children 13 or younger. Thus, excluding companies collected during our web scrape is unnecessary.

3 Experimental Design

To answer our research question, we first need to gather the data necessary for the project. To do this, we utilized a combination of a JavaScript app and a Python program to systematically grab each privacy policy from the apps we were interested in reviewing. From there we segmented each privacy into one to three-sentence chunks using various sentence detection algorithms to create a data format usable by our natural language model. Finally, we analyzed each sentence returned by our text filtering to test whether the chunk complied with any of the three regulations we were scanning for. Described below is a description of the exact methods used in our study.

3.1 Data Collection

To begin such a complex problem, such as classifying compliance via text, we knew that we would need to collect data that could help to train our model down the line. Predictive models depend on labeled data to tell them what a particular text represents. In our case, we needed to construct a model that could identify whether a piece of text from a policy was meeting the specifications in Sections 312.4d(2) or 312.4d(3). This meant that we needed data that took excerpts from actual privacy policies and labeled it saying what section it was fulfilling.

This process was split evenly between all five group members, each taking ten privacy policies. Everyone then read through each policy to pull out any sentences directly correlated with our two desired sections. This data would serve as the backbone for our model so that it could adequately identify different styles of sentences that counted as compliance.

This data was then compiled into a Comma Separated Values file (CSV) to be parsed correctly. Each row of this file contained the company that the privacy policy excerpt was from and the sentence or sentences that correlated with a section. It was then labeled with a zero or one class, corresponding to Sections 312.4d(2) and 312.4d(3), respectively. We chose not to add Section 312.4d(1) to the model as its exact specifications are unsuitable for model prediction. Since this part of COPPA refers to the need to post a specified data collector's name, email, address, and phone number, these pieces of information could better be identified using a parsing method called regular expressions. These are explained further in section 3.4.

3.2 Web Scraping

Due to challenges that arose throughout experimentation, the web scrape portion of our project deviated from the original strategy we planned on implementing. Initially, we had planned to scrape the top 1,000 apps in the "Kids" category of the Google Play Store using a web scraping tool such as Octoparse or Scrapy. However, after researching the best route to complete the web scraping process, we found that we would need to change our goal as the Google Play Store did not offer a way to get the top 1,000 since they only offered apps in categories of 100-200. Additionally, apps within these categories often overlapped, so we needed to remove duplicated applications after completing our initial web scrape. Ultimately, we had just over 500 unique applications from just over 250 unique developers.

We wanted the result of our scrape of the Google Play Store to include the privacy policies for the top 1,000 unique apps in the "Kids" section. Using the "google-play-scraper" tool developed by the user @facundoolano on GitHub, we were able to grab the links of the privacy policies from the apps and to grab each app's full detail, e.g., description, reviews, etc. These full details on each app were nonessential to our experiment, but we kept the information in our result in case it would be helpful for deeper analysis. We

emulated one of this user's methods of retrieval called 'list,' which retrieved a list of apps from a specified collection on the Play Store. Since we only wanted to look at apps from the Kids page, this method suited our experiment best. We implemented features of the 'list' retrieval method, including category (which in our case was family), collection (i.e., top free apps, top non-free apps, and top grossing), the number of apps we wanted to be returned from each collection (which ended up being limited to 200 maximum), and the information we wanted to pull on each application. We then wrote the results of this initial web scrape into an Excel file that could continue to be used.

While our initial web scrape grabbed the links to each application's privacy policy, we additionally wanted to collect these privacy policies to use them in our auditing algorithm. To complete this task, we used the URL library provided by Python to create a script that would visit each privacy policy link and retrieve a copy of the entire web page. Thus, for each app in our Excel file, we appended a column of data that would include the raw content of each privacy policy. While most privacy policies could be written to the file, a few gave us trouble due to the link being inaccessible, the web page blocking our web scraper, the web page being non-decodable, or illegal Excel characters being used by the web page. Adding privacy policies to the result file concluded our scrape of the Google Play Store. It allowed us to begin segmenting the privacy policies to be digested by the auditing algorithm.

While we attempted to gather thousands of applications in our code, the methods we used were limited as they were created in a way to avoid being flagged as bots by Google. From the initial scrape of the Google Play Store, 800 total applications were returned. These 800 applications included several duplicates since we looked for apps across multiple collections, and every app was subject to being contained in just one collection. After filtering the duplicate entries from our file, we were left with 511 unique apps that our scraper returned. Of the 511 apps, 12 privacy policies could not be decoded, 18 privacy policy links could not be accessed, and two apps contained illegal Excel characters. In total, our scrape collected 479 apps' privacy policies from 232 different developers.

3.3 Text Segmentation and Filtering

We decided to use a library called spaCy, an open-source library for Natural Language Processing in Python, to split up a large amount of text from each app's privacy policy. Natural language models typically perform much better when the analyzed text is shorter. Although using spaCy is not necessary for text segmentation, it includes many nice features for the process. We utilized spaCy's load function to connect the prebuilt spaCy pipelines to our Python code. In our case, we installed the English pipeline to help break down sentences. A pipeline is a set definition of various functions and steps that is applied to any piece of text fed to it. These functions transform the text from its original format into sentences.

This step allows our data to be normalized, regardless of its original format. Privacy policies often utilize bullet points and other unusable text display options for the next step in our project. SpaCy can also differentiate parentheses and other odd markings in English and sort them into proper formats. With the pipeline, we can ensure that our model is data in the same, consistent format. Our text segmentation aims to create chunks of three sentences that our model can use to analyze a company's compliance with COPPA. Our program reads each privacy policy into a basic text file before processing it into a spaCy object via the English pipeline. The library can do this via a function called "NLP." This creates chunks of sentences from the processed text.

Once this preprocessing is done, we can extract and embed the sentences into a list. We do this by taking our spaCy processed list of sentences, extracting the first three sentences from the list, and combining the sentences as one large string. This is then stored in a new list within our program. We implement a simple checker to avoid out-of-bounds errors, which would occur when trying to group three sentences if there are less than three remaining in the spaCy object. The system will always attempt to chunk sentences by three until the spaCy processed list ends. For example, if we are at the end of a list with only two remaining sentences, we will need to store the two sentences and not attempt to store a sentence that does not exist because it is the end of the file.

Finally, once a specific policy is parsed and segmented, it is stored in a new data frame containing rows of company names and policy chunks. This data is then utilized during the final step of our experiment.

3.4 Model Creation, Parsing, and Labelling

With the data adequately retrieved and cleaned, we could take each privacy policy and begin classifying to what degree it complied with COPPA standards. This aspect of the experiment utilizes two Python libraries called Tensorflow and BERT. These two packages allow us to construct a model that could identify our two classes, either Section 312.4d(2) or 312.4d(3), based on training data that it received from Section 3.1.

This style of model, created by Google, allows users to take one of their many prebuilt models and apply it to any natural language processing task via fine-tuning and customization. Through this process, a user can provide a model with much fewer data than usual and still see surprisingly accurate classification thanks to the base model's knowledge and definition. From there, the model works by vectorizing the text provided, taking each word, and understanding its position within each sentence, and creating lines of words to derive patterns or meaning. These vectors contain a complex encoding of each word combined with different tokens that help to identify when a sentence begins and ends. This is combined with a transformer called WordPiece Vocabulary. This is a text transformer that is used by BERT models to identify each word or group of words to better maintain critical information by understanding when certain words belong together [7].

Once the model is created, we can begin the fine-tuning process. This is done by providing the model with the data collected in section 3.1. This allows it to understand what sentences should apply to each law section. We can also tell that only two classes should be identified. Other features chosen during the fine-tuning process include batch size, the number of inputs considered before updating the model, and epochs, which are how often the program attempts to retrain the model. Furthermore, validation size, which is what percentage of data should be used to verify whether the model is performing as it should. We also emphasize that the model should focus on the highest accuracy of the validation data, so the model's priority should be to predict the labels in our validation set as accurately as possible [7].

Combining all these choices, we can generate a stable model and predict what class texts are. The BERT library and model train itself multiple times over, constantly trying to improve itself based on previous results. As mentioned above, each epoch means that the program attempts to make a more accurate model than the prior version, focusing on predicting our validation set correctly every time. Once this is accomplished, text can be fed into it, returning a classification.

It is not as simple as saying that a text is either of one section or another. Otherwise, every sentence chunk of a privacy policy would be classified as complying with COPPA. That is why our model returns percentages instead. For each policy segment, it returns two numbers, the percent likelihood that this piece of text meets the requirements of Section 312.4d(2) and the percent likelihood that the text is meeting the requirements of Section 312.4d(3). This then allows us to mathematically determine whether a chunk indeed does qualify as compliance. These percentages can be within the range of zero to one. With sentences, it is more confident in having percentages of 0.75 or greater.

Using these results, we can mathematically determine whether we want to classify text as compliant. Our program requires at least a fifty percent difference between the two classes to be deemed compliant. This is chosen for two reasons. First, chunks that do not meet either class often have roughly equal percentages, meaning that the model cannot tell whether it is one way or the other. This ambiguity translates to a piece of text that meets neither requirement. Second, the requirements of COPPA that we are evaluating are relatively straightforward as to what information needs to be present. Thus, it is logical to conclude that a piece of text that cannot indeed be determined as either class is too vague to qualify as complying with COPPA. This limitation of classification is discussed in section 5.1.

Notably, our model does not interact with the first portion of COPPA we intend to analyze. This is because personal information is often just one instance of the required field in the policy. Thus, it would be impractical to train the model to identify whether

sentences contained this information. This is true because we were interested in how companies comply with Section 312.4d(1). This is because, from our group's own experiences during section 3.1, we discovered that it was prevalent for privacy policies to need one of the four pieces of personal information: name, address, email, and phone number. Thus, we developed a more straightforward approach to identify compliance with this specification.

Utilizing a tool available in Python and other coding languages called regular expressions, we can write specific patterns to be matched that can identify the presence of one of these pieces of information. Using the Python library re, we implemented four regular expressions, each corresponding to one piece of data from Section 314.d(1). These expressions utilize a pattern-matching ability to either optionally look for or definitively target specific characteristics of text that are then returned if a match to our pattern is found. For example, our expression to identify email addresses looks for one or more characters before an at symbol, followed by one or more characters after the symbol. Thanks to the at symbol being used only for emails, we can identify any email contained in a text. This processing style was repeated for name, phone number, and address.

Putting this all together means we can take any privacy policy from a company's Google Play Store page, split it into proper sentences, and identify exact compliance levels on a company-by-company basis. Ultimately, we looked at each company individually to determine overall compliance.

4 Analysis

This section uses the results of our three different scripts combined to evaluate and answer our research questions.

4.1 Presence of Data Collector Information

We addressed **RQ1**, examining how many companies posited the necessary data collector information in their privacy policy. This section of COPPA proved to be the most difficult to prove compliance with as scanning text for the presence of the four pieces of information required by Section 312.4d(1). It was clear right away that our current method of name detection needed to successfully identify names correctly, as most text chunks flagged as names were proper nouns. This aligned with what was found during the data collection outlined in section 3.1. Most companies do not post the name of a specific individual when listing contact information and, instead, typically list the company's name. Thus, our regular expressions were tagging these company names improperly as individual names. This means that we could not accurately state how many companies list or do not list data collectors' names in their privacy policies.

Overall, our collective scrape collected privacy policies from 232 unique companies, giving us a wide variety of data to analyze for compliance. Of these 232 companies, 134 (58%) of them listed an email that could be contacted regarding data issues, 107 (46%) listed a phone number, and 172 (74%) specified where the company was located. Beyond the name issue, our scans for emails, addresses, and phone numbers proved much more successful.

This scanning method for these different pieces could have been better, as seen with the issues regarding name identification. Thus, as before, these percentages are best guesses regarding accurate compliance with these specifications. However, the false positives/negatives rate was low enough to justify including them in the study results. It still needs more compliance in simple requirements such as posting contact and location information. Regarding what is asked of companies, Section 312.4d(1) specifies the least number of requirements. However, only over half of our analyzed companies have this information within their privacy policy as required. This is a disappointing result but comes as no surprise considering the infrequency with which companies violate COPPA requirements. More concerning is that this is the most complied with across our three sections.

4.2 Data Collection Specifics

Regarding **RQ2**, we wanted to see how many companies were directly posting what they are collecting from children 13 years old or younger. This meant devising a way to classify compliance with this stipulation. The tricky thing is that even between two

humans reading the same privacy policy, there will be differences in what is considered "complying" with the law. Thus, there is room for interpretation of the number of compliant companies. Thus, it is essential to acknowledge that our model is trained on what our group determined to meet this requirement.

Furthermore, our process deemed compliance as any piece of text that our model was more than 75% confident that it adhered to our definition of how companies can specify what data is collected. This number was chosen so that only the clearest examples of data collection compliance would be counted and included. This value was chosen as it did not make sense to classify any piece of text with over a 50% confidence value as compliant. Since the model returns an approximately 50-50 split if unsure what class a piece of text is, we needed a cut-off to say whether a text met the compliance goals.

Using our confidence definition, we found various examples of text that met our definition of complying with Section 312.4d(2). An example of such identification is:

"Our games do not collect any personal information and do not share it with third parties. Hence, they are COPPA-compliant. Our games do not require extra permissions, so we cannot collect personal data and do not want to."

This text is just one of many examples of privacy policy components that met our model's definition of compliance.

Considering all the classified text segments, we found that only 91 (39%) of our scanned companies had at least one example of a text chunk that specified what data was being collected and was thus in compliance with section 312.4d(2) of COPPA. This figure is significantly lower than the percentages found in section 4.1 were incredibly disheartening. This section is the most important concerning protecting children and informing their guardians as to what their kids are interacting with. To discover that only a little over a third of our companies included clear enough language to qualify as this could have been better.

4.3 Right to Review and Delete

Finally, our research aimed to address **RQ3**, asking what portion of our analyzed companies tried to include information and instructions on how parents and guardians could review, update, and request the deletion of any data collected from their children. Like section 4.2, this method of compliance analysis was heavily dependent on how we classified compliance during data collection. The exact 75% confidence estimate was used to determine whether a chunk of text was classified as having complied with section 312.4d(3).

As an example, the following text was classified as having complied with Section 312.4d(3):
"We ensure we do not store your information for a longer period than necessary basis of processing. When we collect, use, store or process, in any other way, your information, we rely on a number of legal bases, as set forth in this Privacy Policy: Consent: we rely on your consent to store and use your personal information you provided to us You may withdraw your consent at any time by contacting us at . If you do not consent to the use of your personal data, we may not be able to provide you with all or parts of our services."
As seen above, this text mentions that you may withdraw your consent anytime. This meets our group's definition of compliance with Section 312.4d(3).

After our predictions were generated, we found that only 61 (26%) companies in our dataset had included text that met our model's definition of compliance with this section. To see a figure this low was not shocking. During data collection, we found that it was much less common for companies to include text that met our group's definition of compliance. Thus, discovering that very few companies meet this requirement aligns with what was seen earlier.

This is, nonetheless, the section with the least overall adoption and, thus, an area with the most room to grow.

5 Discussion

In this study, we collected the privacy policies of almost five hundred different apps that are available

and marketed to children on the Google Play Store to determine to what degree they comply with the specifications of Section 312.4 of COPPA. In summary, we found that:

-**RQ1:** We could not determine compliance with including the data collector's name. It was found that 134 (58%) of companies posted their email address, 107 (46%) specified a phone number, and 172 (74%) listed the address of their company itself. Section 312.4d(1) was the most complied with a portion of COPPA that was studied.

-**RQ2:** Only 82 (35%) of companies had at least a single portion of their privacy policy that was more than 75% likely to specify what type of data was being collected of children by their application. This means that section 312.4d(2) was the second most complied with aspect of our study.

-**RQ3:** It was found that 91 (39%) of analyzed developers included information on how to contact them to review, request, and delete personal information collected from children. This meant that section 312.4d(3) of COPPA was complied with the least of the three.

Our research has shown that there is still a long way to go until companies fully comply with this specific portion of COPPA. We found only 24 (10%) examples of companies that had met all the possible compliance requirements, a paltry number compared to the scope of our data. While over half of the scanned policies contained at least one piece of information that addressed data collector information, it is incredibly disappointing to see companies not listing the correct information regarding data collection and review.

This study, however, could be more extensive and is only scratching the surface regarding being able to say how many companies are complying with section 312.4d confidently. Our methods, while logical in theory, would require a much more extensive collection of training data and a more precise method of text scraping and segmentation that was not found during our research.

5.1 Limitations

One of the most significant limitations of this experiment is the human element. By human nature, each group member would determine what text counted as complying with COPPA requirements differ from one another. This conflicting perception can only be mitigated so much by the discussion we had when collecting data during the steps described in section 3.1. Thus, since it is nearly impossible to get humans to agree on what explicitly counts as meeting the requirements, it stands to reason that a computer model will suffer the same fate since it is trained on the personal bias that each of us introduced. This means that despite our best efforts, without a more rigorously developed and more extensive training set, our model is limited to only being as accurate as we made it via the training data we collected.

If this study were to be repeated, starting with a much more extensive training data set for the BERT model would be incredibly beneficial. Without it, our model is left with a much larger margin for error and misclassification despite our best efforts, as it is only aware of a limited number of examples of what is considered as meeting compliance with our various sections of COPPA. Alongside this, there would also need to be a significant increase in computing power for the scripts. Currently, our final script that operates after the web scrape is completed takes approximately two hours to run on a single PC. An increase in data on both the training and analysis sides would significantly increase this figure.

Another aspect that holds back our ability to be entirely conclusive is identifying proper names to identify the presence of a data collector's name in a policy. As mentioned in section 4.1, we had to remove this classification from our results as there were too many false positives using our current method. Future repetitions of our study need to implement a more robust and precise method of scanning for names, as utilizing regular expressions does not enable proper classification.

Mainly, the two challenges faced by our web scrape were that some policies would need to be better defined within the website's HTML or some other strange exception that made it impossible to grab the privacy policy from that app's website despite our best efforts.

Secondly, our web crawl could only be so effective due to various issues on certain company pages that prevented us from including it in the dataset despite our best efforts. Secondly, it was common for us to discover that certain websites blocked our crawling script entirely. This, too, was not surprising, as many pages have restrictions on non-human access to be more protective over who or what is accessing their data, ironically. These two factors meant that we had fewer companies overall to evaluate our results on, which could mean that our findings could show a higher or lower compliance rate than what was calculated.

Another limitation of this project was that it needed to be very minimal in scope to be feasible. COPPA is a wide-ranging law that contains many different regulations that could be tested for and evaluated. Section 312.4d appeared to us as the most straightforward to evaluate, thanks to how it clearly stated what information must be contained in a company's privacy policy. We also needed to limit the study to only one type of child-marketed online service, as there are hundreds of avenues for children to interact with different websites and services that could collect their data. Future works could likely build off what we have created to investigate compliance with this section of COPPA further or be applied to a different medium of children's web access, such as websites or Internet of Things devices. Focusing on apps from the Google Play Store gave us a clear goal as to what we wanted to investigate and a baseline understanding that all the data we were hoping to collect would be available all in the same place.

5.2 Recommendations

The main point that can be taken from this study is that we must know what companies are complying with federal regulations and which ones need to be improved. Technology is ever evolving, and we must stay on the offensive to prevent issues instead of dealing with them once it becomes too late. Thanks to our proof of concept, it would be possible for federal regulators to create a process like our methods to better check for compliance. With a larger, more refined dataset and a more consistent definition of compliance, the company's policies could be screened as soon as they are posted to a service like the Google Play Store.

This is only one piece of the compliance puzzle. Our methodology only tests for the presence of writing that states what the company is collecting, whom to contact if there are issues, and what the process is due to review and delete data. We do not test in any way what companies are doing behind the scenes. This would require a different style that cross-references what is stated to what is happening. These two studies, in tandem, would then be able to determine actual compliance with what COPPA specifies is required of companies.

We hope, too, that this verification style can continue to grow alongside technology so that COPPA can continue to make a difference in protecting children's online safety. We also desire to see further updates be made to COPPA's contents so that it can keep up with an increasing number of new avenues of data collection. Expressly, sections regarding mobile devices acknowledge that most children's online interaction occurs when there is no direct parental supervision, as envisioned during COPPA's original passing. With this being the twenty-fifth anniversary of the original passage of COPPA and ten years since the last comprehensive update, lawmakers must review what is currently covered in the law and amend it to meet the growing needs of consumers across the United States.

REFERENCES

[1] FEDERAL TRADE COMMISSION, 2020. Complying with COPPA: Frequently Asked Questions. https://www.ftc.gov/business-guidance/resources/complying-coppa-frequently-asked-questions

[2] CODE OF FEDERAL REGULATIONS, 2013. Part 312 – Children's Online Privacy Protection Rule. https://www.ecfr.gov/current/title-16/chapter-I/subchapter-C/part-312

[3] O'MELVENY, 2023. FTC Obtains Record Penalties from Video Game Company Amidst Growing Privacy and Consumer Protection Enforcement Trends. http://www.omm.com/resources/alerts-and-publications/alerts/ftc-obtains-record-penalties-from-video-game-company/#:~:text=The%20US%24275%20million%20penalty,affirmative%20consent%20from%20their%20parents.

[4] REYES, I, ET AL. 2018. "Won't Somebody Think of the Children?" Examining Coppa Compliance at Scale. *Proceedings on Privacy Enhancing Technologies vol. 2018 no. 3* 63-83.

[5] BREAUX, T.D., ET AL. 2021. Analyzing privacy policies through syntax-driven semantic analysis of information types. *Information Software and Technology vol 138*

[6] VAN NORTWICK, MAGGIE, WILSON, CHRISTO. 2022. Setting the Bar Low: Are Websites Complying with the Minimum Requirements of the CCPA? *Proceedings on Privacy Enhancing Technologies, 2022* (1), 608-628.

[7] SHEKAR, CHANDRA. 2022. *Simple Text Multi Classification Task Using Keras BERT.* https://www.analyticsvidhya.com/blog/2020/10/simple-text-multi-classification-task-using-keras-bert/

Appendix

Our project was developed simultaneously by five students, with key components of the project being split amongst each member. Writing of the report and collection of data for the BERT model was accomplished on an equal basis, while the programming was split into three primary categories. Evan and Lydia worked on the web scrape, Sebastian and Mark provided text cleaning and segmentation, and Peter created the BERT model, regexes, and combined the code into one file. The code for the project can be found at the following GitHub link: https://github.com/lesbaum/Privacy-Law-Technology-Project