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, regulation 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.

In order to comply with COPPA's regulations, three things must be displayed on a website regarding the data collection of children:

1. 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.
2. There must be a clear description of what kinds of data are being collected from users who are thirteen years old or younger.
3. 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.

In this study, we investigate how app developers on the Google Play Store with apps explicitly directed at children comply with COPPA's notification requirements. Utilizing a one-time web crawl, we collected the privacy policies of over five hundred apps found in the "Kids" section of the Google Play Store. Using these policies, we can analyze the content to see which portions of notifications are included and to what extent. Using our results, we can better understand just 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 an 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, addresses, email, and phone numbers).
2. A section stating what is being collected.
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 or not 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.

In order 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. It is clear that 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 different aspects of COPPA.

**2.1 Prior Work**

The Children's Online Privacy Protection Act (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 general 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. Similar to 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 only we will 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 hope to do 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 to this 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 or not 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 or not 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, in order 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. As long as 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 or not 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**

In order 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**

In order 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 or not a piece of text from a policy was meeting the specifications in either section 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. Each individual 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 a top 1,000. We instead only offered categories that could be scraped in groups of 100-200 applications. 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 in our code to gather thousands of applications, 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 pretty 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 in order 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 that are chosen during the fine-tuning process include batch size, which is the number of inputs considered before updating the model, epochs, which are how many times the program attempts to retrain the model, and validation size, which is what percentage of data should be used to verify whether or not 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 of 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 and can return 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 or not 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 determine mathematically whether or not we want to actually classify text as compliant or not. Our program requires at least a fifty percent difference between the two classes in order to be deemed compliant. This is chosen for two reasons. First, chunks that don't meet either class often have percentages that are roughly equal to each other, 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 is needing to be present, thus it is logical to conclude that a piece of text that cannot truly be determined as either class is too vague to possibly qualify as being in compliance with COPPA. This limitation of classification is discussed in section 5.1.

Notably, our model does not interact with the first portion of COPPA that we are intending to analyze. This is because personal information is often just one instance of the required field located somewhere in the policy, thus it would be impractical to train the model to identify whether or not sentences contained this information. This is further true because we were interested to see to what extent companies are complying with Section 312.4d(1). This is because from our group's own experiences during section 3.1, we discovered that it was incredibly common for privacy policies to be missing one of the four pieces of personal information: name, address, email, and phone number. Thus, we developed a simpler 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 different 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 certain 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 piece of text. This style of process was repeated for name, phone number and address.

Putting this all together means that 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. In the end, we looked at each company one at a time to determine overall compliance.

**4 Analysis**

In this section, we use the results of our three different scripts combined to evaluate and answer our research questions.

**4.1 Presence of Data Collector Information**

We began by addressing **RQ1**, looking at how many companies are positing 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 different pieces of information required by Section 312.4d(1). It was clear right away that our current method of name detection was not successfully identifying names properly, as most text chunks that were flagged as names were actually 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 specific individual when listing contact information, and instead typically list the name of the company itself. Thus, our regular expressions were tagging these company names improperly as individual names. This means that it was not possible for us to state with any degree of certainty how many companies list or do not list data collector’s names in their privacy policies.

Beyond the issue with names, our scans for emails, address, and phone numbers proved to be much more successful. 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.

This method of scanning for these different pieces was not perfect, as seen with the issues regarding name identification. Thus, as before, these percentages are best guesses in regards to true compliance with these specifications, but we felt that the rate of false positives/negatives was low enough to justify including them in the results of the study. It is still nonetheless disappointing to see so little compliance in such a simple requirement such as posting contact and location information. In terms of what is asked of companies, Section 312.4d(1) specifies the least number of requirements, and yet only a little over half of our analyzed companies have this information contained within their privacy policy as required of them. This is a disappointing result to find but comes as no surprise considering the infrequency in which companies are found to be in violation of COPPA requirements. More concerning is that this is the section that is most complied with across our three different sections.

**4.2 Data Collection Specifics**

In regard to **RQ2**, we wanted to see in a concrete way how many companies were directly posting what it is they are collecting from children who are 13 years old or younger. This meant devising a way to classify compliance with this stipulation. The tricky thing regarding this is that even between two humans reading the same privacy policy, there will be differences in what is considered as “complying” with the law. Thus it is important to acknowledge that our model is trained on what our group determined to be meeting this requirement, and thus there is room for interpretation of the number of compliant companies.

Furthermore, our process deemed compliance as any piece of text that our model was greater than 75% confident that it did adhere 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 we felt that it did not make sense to classify any piece of text with over a 50% confidence value as a actually being compliant. Since the model returns an approximately 50-50 split if it is unsure of what class a piece of text is, we needed a cut off to say whether a text met the goals of compliance.

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

“Our games don't collect any personal information and doesn't share it with third parties Hence they are COPPA-compliant Our games don't require any extra permissions so we can't collect any personal data and don't want to”

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

Taking all of the classified text segments into consideration, 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 being significantly lower than the percentages found in section 4.1 were incredibly disheartening. This section is one that we felt is the most important in regard to 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 was not ideal.

**4.3 Right to Review and Delete**

Finally, our research aimed to address **RQ3**, asking what portion of our analyzed companies made an effort to include information and instructions on how parents and guardians could review, update, and/or request the deletion of any data collected of their child. Similar to section 4.2, this method of compliance analysis was heavily dependent on how we classified compliance during data collection. The same 75% confidence estimate was used to determine whether or not a chunk of text 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 make sure 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 [email&#160;protected]. 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 does include a mention of directly that you may withdraw your consent at any time. 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 total 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 are meeting this requirement aligns with what was seen earlier.

This is nonetheless the section with the least overall adoption, and thus an area where there is 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 in order to determine to what degree they are in compliance with the specifications of Section 312.4 of COPPA. In summary, we found that:

**-RQ1:** It was not possible for us to truly determine compliance with the inclusion of 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 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 be specifying 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 as to how to contact them to review, request, and delete personal information collected of children. This meant that section 312.4d(3) of COPPA was complied with the least of the three.

Overall, our research has shown that there is still a long way to go until companies are truly fully compliant with this specific portion of COPPA. We found only 24 (10%) examples of companies who had met all the possible aspects of compliance, 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 proper information regarding data collection and review.

This study, however, is incredibly limited in nature, and is only scratching the surface in terms of being able to confidently say just how many companies are complying with section 312.4d. Our methods, while logical in theory, would require a much larger collection of training data, as well as a clearer method of text scraping and segmentation that was not found during our research.

**5.1 Limitations**

One of the largest limitations of this experiment is the human element. By virtue of human nature, each member of our group was going to determine what text counted as being in compliance with COPPA requirements differently 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 larger 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, it would be greatly beneficial to start with a much larger training data set for the BERT model. 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 large increase in the amount of computing power for the scripts themselves. As of current, our final script that operates after the web scrape is completed takes approximately two hours to run on a singular PC. An increase in data on both the training and analysis sides would greatly increase this figure.

Another aspect that holds back our ability to be fully conclusive is that of identifying proper names in order 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 final results as there were simply too many false positives using our current method. Future repetitions of our study would need to implement a more robust and clear method of scanning for names, as utilizing regular expressions does not enable proper classification.

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. Mainly, the two challenges faced by our web scrape was that some policies would be so poorly defined either within the HTML of the website itself or some other strange exception that it made it impossible to grab the privacy policy from that app’s website despite our best efforts. Secondly, it was not uncommon for us to discover that certain websites were blocking 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 overall companies to evaluate our results on, which could mean that our findings could show either a higher or lower compliance rate than what was calculated.

Another limitation to this project was that it needed to be very minimal in scope in order 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 needed to be contained in a company’s privacy policy. It was also necessary that we limit the study to only one type of child marketed online service, as there exists hundreds of avenues for children to interact with different websites and services that could be collecting their data. By focusing on apps from the Google Play Store, it gave us a clear goal as to what we wanted to investigate, as well as a baseline understanding that all the data we were hoping to collect would be available all in the same place. It is very possible that future works could build off what we have created to either further investigate compliance with this section of COPPA or be applied to a different medium of children’s web access, such as websites, or recently internet of things devices.

**5.2 Recommendations**

The main point that can be taken from this study is that it is vital that we know what companies are actually complying with federal regulations and which ones are lacking. Technology is ever evolving, and it is vital that we stay on the offensive to prevent issues instead of dealing with them once it becomes too late. Thanks to our proof of concept, we believe that it would be possible for federal regulators to create a process similar to our methods in order to better check for compliance. With a larger, more refined dataset, and a more consistent definition of compliance, it is very possible that company’s policies could be screened as soon as they are posted to a service such as 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, who to contact if there is issues, and what the process is due to review and delete data. We do not test in any way for what companies are doing behind the scenes. This would require a different sort of style that would cross reference what is stated to what is actually happening. These two studies in tandem would then be able to determine true compliance with what COPPA specifies is required of companies.

We hope too that this style of verification can continue to grow alongside technology, so that COPPA can continue to make a difference in protecting the online safety of children. 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. Specifically, sections regarding mobile devices that acknowledge the fact that most children’s online interaction occurs when there is not direct parental supervision, as was 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, it is critical that lawmakers 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/

*Conference Name:ACM Woodstock conference*

**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, Evan 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>

*Conference Short Name:WOODSTOCK’18*

*Conference Location:El Paso, Texas USA*

*ISBN:978-1-4503-0000-0/18/06*

*Year:2018*

*Date:June*

*Copyright Year:2018*

*Copyright Statement:rightsretained*

*DOI:10.1145/1234567890*

*RRH: F. Surname et al.*

*Price:$15.00*