COMM 493 Final Report

Cyber Shield

**Submitted by: Team 5**

Hardik Uppal

Shubham Barnwal

Kristen Perry

Stefan Negus

Application URL: <http://hardikuppal.pythonanywhere.com/>

GitHub Code Link: <https://github.com/hardik-uppal/cyber_shield>

**Section 1: The Business Case**

**1.1 Problem**

Social media has drastically changed the way in which we interact and communicate. The rapid growth and adoption of social media, coupled by the continually changing technologies that host these social media platforms means that these platforms are constantly evolving. However, as social media continues to change, our understanding of the dangers associated with using these applications becomes clouded. Specifically, due to the differences of understanding these applications between generations, youth are particularly susceptible to danger when using social media. Parents of young children using social media, like Instagram and Snapchat, typically do not understand these applications because they did not grow up using social media themselves. Cyberbullying, a form of harassment through electronic means, occurs on social media and places children at risk for harassment, often leading to poor mental health manifestations. The content of these applications are monitored loosely, meaning that unsupervised activity may lead to your child interacting with inappropriate or explicit content. Moreover, very few parental controls exist on these applications, particularly Instagram. Even if parents are able to directly monitor their children on these applications, they typically don’t have the time to read through every post, comment or tag to search for dangerous content. Parents need a fast, effective and accessible way to monitor their child’s social media activity to ensure their child is staying safe online.

**1.2 Value Proposition & Solution**

Our application allows you to protect your child's safety online by monitoring their social media activity for cyber bullying and mental health concerns, saving you time and putting your mind at ease.

Our MVP targets your child’s Instagram activity. Instagram was chosen as our primary application due to its widespread use by youth, the lack of parental controls, explicit content and the availability of its data for analysis. By logging into our application with your child’s credentials, a summary report is instantly generated that displays basic information about your child’s account, such as total number of posts and number of explicit comments identified. Dangerous comments have been classified into five categories: identity hate, obscene, threat, toxic and neutral. A table displaying the 5 users who have posted the most dangerous comments is displayed, so parents can understand if one particular user is targeting their child. The next table, comment list, displayed each dangerous comment, the user who posted this comment, the category of the comment, date and time of posting. This table also leverages tone analysis to display the tone of the comment, offering further insight for the parent. A third table breaks down the number of dangerous comments by day, allowing parents to understand the time period in which their child experiences the most dangerous comments. Finally, parents have the option to individually analyze dangerous comments, or enter their own comments to assess how each comment is classified.

Beyond the main page of the application, an FAQ page has been developed that includes general application information, and a contact us page was included for additional questions or concerns.

**1.3 Primary Persona**

The target market for Cyber Shield is primarily consists of middle-aged parents with children between the ages of eight and sixteen. These families live in urban or metropolitan areas and are generally unfamiliar with social media and technology, but may have a simple Facebook or Instagram account they rarely use. They also work mainly full-time, and do not have time to continually monitor their children’s social media activity. A detailed primary persona can be found in the appendix.

**1.4 Competitor Landscape**

Overall, Cyber Shield is competing with various parental control applications. Some commonly used applications include Kidgy, mSpy, ScureTeen Parental Control and ESET Parental Control. These apps simply monitor online internet browsing, and allow you to lock and unlock some apps for use by your children. Some also allow you to limit the amount of time or times of day that children can use these applications. However, these apps pull absolutely no insight into what your child is interacting with online, it can only limit their activity. Therefore, Cyber Shield differentiates in the apps focus on providing actionable insight for parents based on comments.

Specifically, one application called Bark focuses on tracking all activity on a child’s phone and provides notifications and insight when dangerous activity is detected. Bark tracks content 24/7 across all social media platforms, text messages, email monitoring and phone calls. The app covers indications of cyberbullying, suicide and depression, online predators, adult content and more. Parents can have an unlimited accounts for an unlimited number of apps. However, while Bark monitors videos and images on Instagram, it does not monitor comments and direct messages. Moving forward, we would need to differentiate ourselves from this large competitor. While Bark provides an almost invasive holistic view, the parent would only get a notification when dangerous activity is detected, but very minimal context as to what caused this alert. With Cyber Shield, you have the ability to see the comment detected immediately, the user who posted this comment and when it was posted.

**1.5 Business Model**

Cyber Shield will us a subscription-based model with fixed monthly payments. While the website and future mobile application will be free to access and download, the parent will have to purchase an account to view their dashboard. Benchmarking against similar applications previously mentioned, all websites and applications in the industry leverage a subscription model, some with add-on paid features. Our closest competitor, Bark charges $9 per month for unlimited use for a family. Cyber Shield will first first charge $5 per month for unlimited usage per family. In the future once we have added additional social media platforms, this price may increase. We will also offer a two week free trial for parents, allowing them to get comfortable with the app and determine if they want to sign-on for monthly payments following the trial.

**Section 2: Technical Solution**

**2.1 MVP Status**

The application currently fetches live comments from Instagram and classifies it into 5 categories using a Bluemix NLC classifier which is trained on our toxic comment dataset. We also analyzed the tone of the comments in order to improve classification accuracy using the IBM Bluemix Tone analyser. This helps to specify whether the comment was related to cyberbullying, which is mostly associated with ‘anger’ tone classification. We are currently using the Instagram API in a sandbox environment, which only captures comments of a few registered users along with the user of account that is being analyzed. These comments are then stored in local Sqlite database and are populated using Python and JavaScript on an HTML dashboard page.

Our application includes a backend script which is scheduled to continuously monitor new comments on the user’s Instagram handle and a web application which helps in visualizing inferences from generated metadata.

**2.2 Application Performance & Metrics**

Performance Metric Table of NLC:

|  |  |  |
| --- | --- | --- |
|  | **TRUE POSITIVE** | **TRUE NEGATIVES** |
| **PREDICTED POSITIVES** | 11 | 2 |
| **PREDICTED NEGATIVES** | 1 | 2 |

Performance Metric Table of Tone Analyzer:

|  |  |  |
| --- | --- | --- |
|  | **TRUE POSITIVE** | **TRUE NEGATIVES** |
| **PREDICTED POSITIVES** | 8 | 0 |
| **PREDICTED NEGATIVES** | 5 | 3 |

**2.3 Data Used**

The data that was used in Cyber Shield’s application is a data set from Kaggle, using Wikipedia comments to classify into five main categories. The comments were flagged as either toxic, obscene, threat, identity hate, or neutral. A neutral comment is a general comment that is not considered negative. Identity hate is a comment that attacks a person or group on the basis of attributes such as race, religion, sex, etc. A threat is a statement of intention to inflict pain, injury, damage, etc. to someone else. An obscene comment is an extremely offensive word or expression. Finally, a toxic comment is a rude, disrespectful or unreasonable comment that is likely to make you leave a discussion.

**2.4 Design Flow**

**2.5 Future Iterations**

Moving forward, we have brainstormed multiple features and extensions to expand the functionality of our application. First, visual recognition to analyze and assess the actual photos that children post onto their accounts is an immediate concern we would like to address. Allowing visual recognition of photos would essentially expand the breadth of our app and further recognize dangerous activity that may put children at risk. Next, we hope to include a video tutorial explaining how our software operates and how parents can best use our application. As well, include a list of parenting resources so that when danger is detected, parents are informed on how to deal with difficult situations. Next, push notifications that immediately alert parents when dangerous comments or posts are detected would allow parents to act quickly when their child is at risk. Additionally, we hope to expand our application to allow multiple accounts to be monitored. We would also enable SSL security (HTTPS) on the web application to make sure that there is a secure and encrypted connection between the server and browser.

We also hope to develop monthly reports that allow parents to quickly review their child’s online activity. Additionally, we would like to develop a page on our application that informs children how their login information may be used. We recognize a weakness of our application is that we require the child’s password to access their account; this page would inform the child that the parent has no control over their child’s account and cannot access their personal messages. We hope to improve our performance metrics through allowing users to identify when comments are misclassified, improving the performance of our NLC.

After these smaller iterations, the development of a mobile version of our application would allow this information to be more readily available to parents. Finally, we hope to expand our application beyond Instagram, so that parents may be able to monitor their children’s social media activity through our app on other social media platforms.

**2.6 Strengths of Technical Solution**

* The application enables parents to monitors their child’s account for cyberbullying comments on posts and take measures if needed.
* The application only reads comments on Instagram posts and does not post or delete comments, giving the owner of the account more control.
* It reports only explicit comments and not other comments, keeping check on cyberbullying and giving the account owner their space and privacy.

**2.7 Weaknesses of Technical Solution**

* The current MVP is just for public comments posted on the Instagram, direct message and images/videos are currently not a part of the application.
* The NLC classifier (lite version) we used took only 20,000 rows off the labeled data for training. Its accuracy could have been better if we had been able to process all the 150,000 rows.
* To access the account through Instagram’s API we would need login information of the account owner every time the “access token” expires.
* Since we classify the data while loading it to the database, it makes the first rendering of the page a little slow.

**Appendix**

A sample customer persona:

