

Watchdog

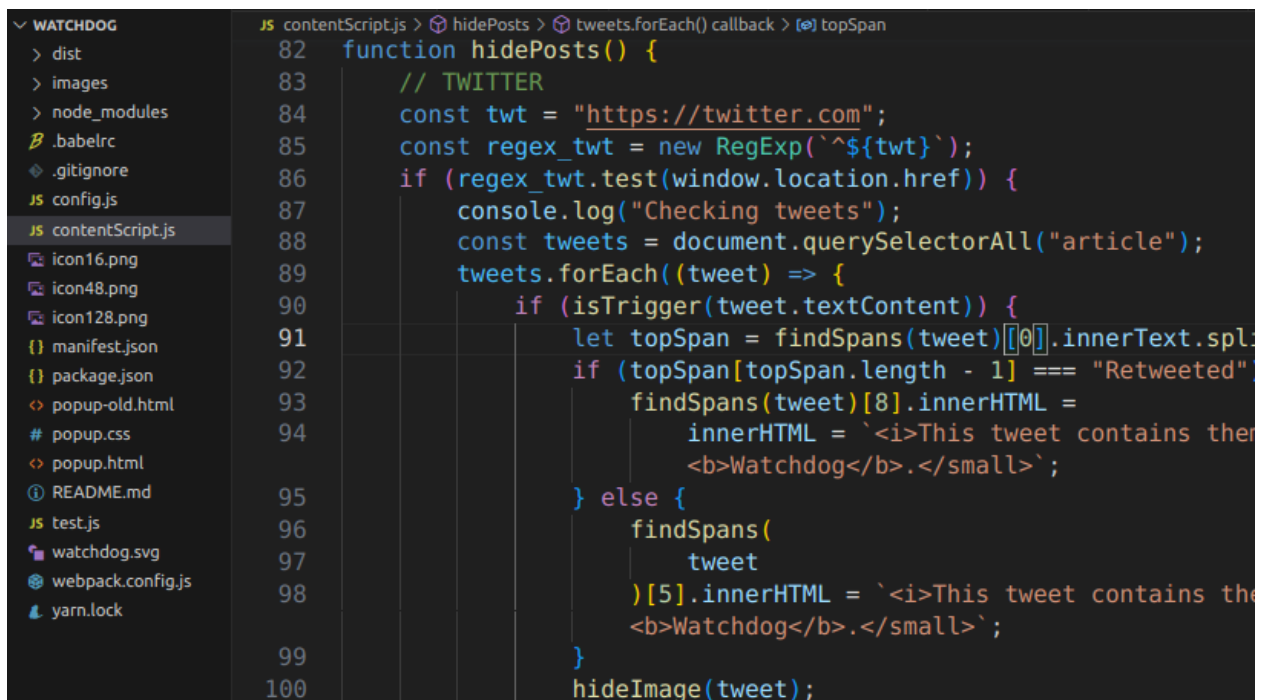
Proof of Concept

Introduction

Watchdog is a browser extension on Google Chrome that protects the user from encountering potentially triggering posts while they scroll through their feed. If the extension detects a potentially triggering post, it immediately replaces the text with a message explaining why it's blocked and blurs out any accompanying images. This document covers the features in the base version of the submitted prototype.

Technical Implementation

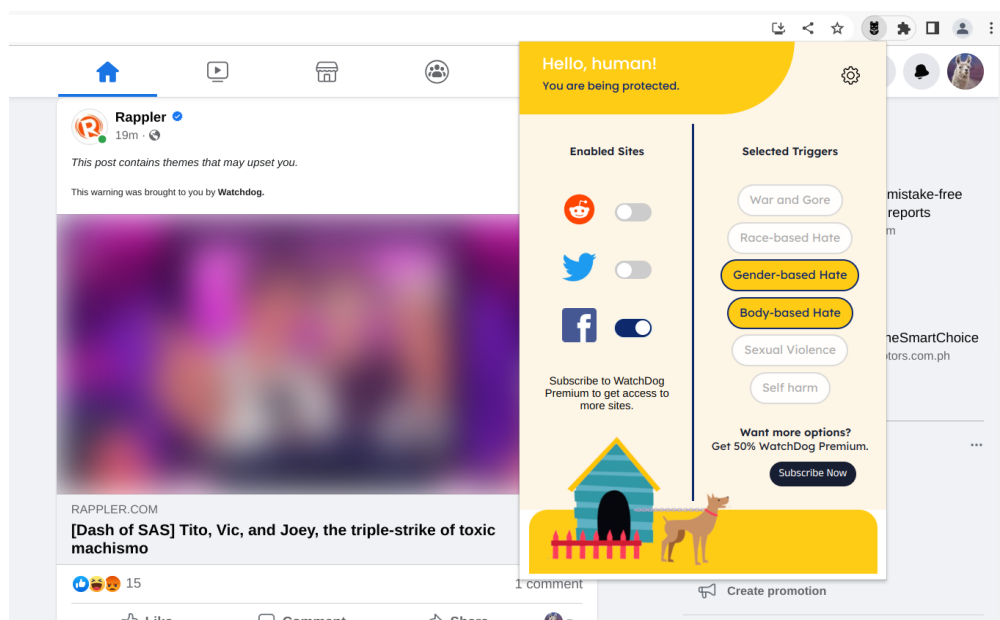
See the full codebase in the repository [here](#). Much of the flow can be found inside `contentScript.js`. The extension reads each post using a `querySelector` and makes a call to OpenAI, which determines whether or not the post is potentially triggering, based on the user's saved preferences. Posts are censored accordingly, using DOM manipulation.



```
JS contentScript.js > hidePosts > tweets.forEach() callback > topSpan
82 function hidePosts() {
83   // TWITTER
84   const twt = "https://twitter.com";
85   const regex_twt = new RegExp(`^${twt}`);
86   if (regex_twt.test(window.location.href)) {
87     console.log("Checking tweets");
88     const tweets = document.querySelectorAll("article");
89     tweets.forEach((tweet) => {
90       if (isTrigger(tweet.textContent)) {
91         let topSpan = findSpans(tweet)[0].innerText.split(" ");
92         if (topSpan[topSpan.length - 1] === "Retweeted") {
93           findSpans(tweet)[8].innerHTML =
94             innerHTML = `This tweet contains the`
95             <b>Watchdog</b>.</small>`;
96         } else {
97           findSpans(
98             tweet
99           )[5].innerHTML = `This tweet contains the`
100             <b>Watchdog</b>.</small>`;
101         }
102       }
103       hideImage(tweet);
104     });
105   }
106 }
```

```
// Call GPT
const chat = `Answer only with 'Yes.' or 'No.'. Does the following tweet cover one of the following trigger categories: ${triggers}? "${text}"`;
const params = {
  messages: [
    {
      role: "system",
      content: "You are an application that determines whether the given text might be triggering.",
    },
    {
      role: "user",
      content: chat,
    },
  ],
  model: "gpt-3.5-turbo-16k-0613",
  max_tokens: chat.split(" ").length,
  temperature: 0,
};
client
  .post("https://api.openai.com/v1/chat/completions", params)
  .then((response) => {
```

Each call to OpenAI takes the user’s chosen set of triggers and the text within a particular tweet or post. The prompt is then analyzed by GPT, who is instructed only to reply with “Yes.” or “No.”, limiting the maximum number of tokens to two, for efficiency’s sake. As accuracy is vital, the randomness (or “temperature”) is set to zero.



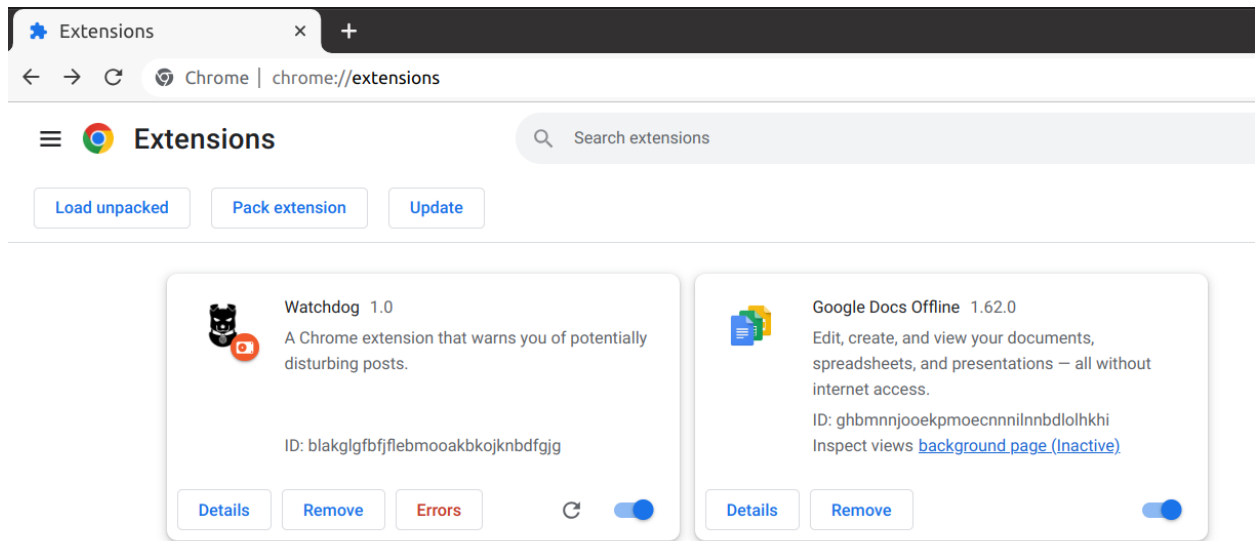
User preferences are stored using the Chrome storage API. The user may change their preferences with regard to which sites should be monitored and which categories should be enabled by clicking on the popup when the extension is clicked from the browser’s navbar.

Finally, the files are bundled and exported using webpack. Configuration keys are stored in an obfuscated file called config.js.

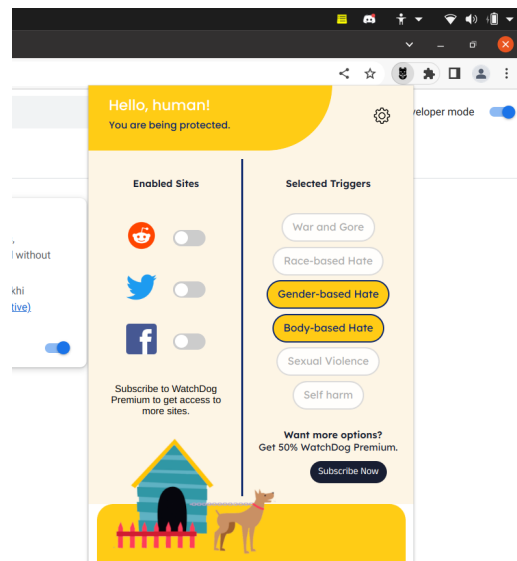
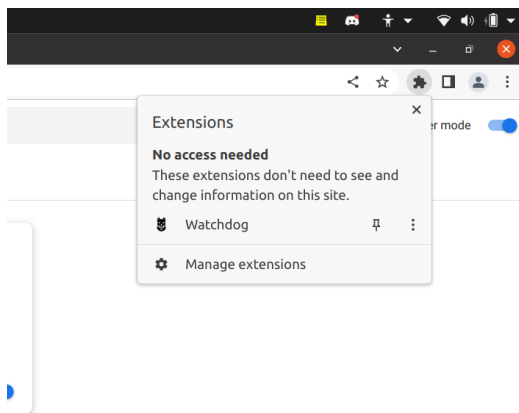
Running the Code

Running and using the program is simple. All the user needs to do is the following:

1. Download the release from <https://github.com/ReneXiong/Watchdog/releases/tag/v1>.
2. Unzip the file.
3. Open Google Chrome and enter `chrome://extensions` in the address bar.
4. Click “Load Unpacked” and select the unzipped folder “watchdog”.



To modify their preferences, the user may click on the puzzle icon in Chrome’s navbar and click on the Watchdog icon to view the popup.



Input and Expected Output

The extension is coded only to run on Twitter and Facebook. Further, the extension will only work on web browsers as the program expects a certain DOM structure for each site. The extension is then expected to discern whether or not each post is a trigger. If it is a trigger, the post's text will be replaced with a message and any image or video attached to it will be blurred.

←

BBC News (UK) 
525.5K Tweets

Follow



BBC News (UK)  @BBCNews · 38m
'The Cuban regime killed my father' - dissident's daughter

...



←

BBC News (UK) 
525.5K Tweets



BBC News (UK)  @BBCNews · 38m
This tweet contains themes that may upset you.

This warning was brought to you by **Watchdog**.



Follow

Hello, human!
You are being protected.

Enabled Sites

 ☐

 ☒

 ☐

Subscribe to WatchDog Premium to get access to more sites.

Selected Triggers

☒ War and Gore

☐ Race-based Hate

☐ Gender-based Hate

☐ Body-based Hate

☐ Sexual Violence

☐ Self harm

Testing

```
• boberoo@boberoo:~/Desktop/Watchdog/extension/watchdog$ node test.js
The young man jumped off the building to kill himself
Is this triggering?
Yes.
Code execution time: 868.792338013649 milliseconds
• boberoo@boberoo:~/Desktop/Watchdog/extension/watchdog$ node test.js
Albert Camus once wrote that suicide is philosophically absurd
Is this triggering?
No.
Code execution time: 889.1632729768753 milliseconds
• boberoo@boberoo:~/Desktop/Watchdog/extension/watchdog$
```

We created unit tests of various prompts for each category using a [unit test file](#). After ten test cases the average came to about 1059.1824 milliseconds, or 1.059 seconds. This will certainly be less time than it takes a user to read and comprehend a tweet.

We also looked at the program's ability to filter reliably by testing across different prompts. For instance, we compared texts that used the same word in different contexts. Certain contexts, such as news reports, were identified as triggers and other contexts, such as abstract literature, did not. We tested the extension across twelve different prompts (two for each trigger category) and found that OpenAI was able to discern the context for all of them.

Limitations and Future Enhancements

The main limitation our application faces is the limited amount of control over third-party APIs such as OpenAI. It's crucial that we are able to develop an in-house language model moving forward so that we are able to improve trigger recognition. Such a system should allow Watchdog to create a smarter and more personalized user experience. For instance, this should allow the user to "report" posts that the extension was not able to identify and filter and take such an event into account in the future. This should also allow for more inclusive experiences for users who have very specific triggers that are not immediately obvious.

References

Wenz, J. (2015, October 27). How many characters make up a tweet? Ask the experts. *Wired*. Retrieved from <https://www.wired.com/2015/10/many-characters-tweet-ask-experts/>

OpenAI documentation. (n.d.). API Reference. OpenAI. Retrieved from <https://platform.openai.com/docs/api-reference>