

Advertisement detection system using machine learning techniques.

Final Project Report

DT211c

BSc in Computer Science (Infrastructure)

**Jack Duggan**

**C16350866**

**Brian Gillespie**

School of Computer Science

Technological University, Dublin

**Date**

Abstract

This project revolves around the detection of advertisements using machine learning means, creating a system that can apply these means and the evaluation of that system. The motivation behind this system is the turmoil in advertisement deliverance. The use of advertisements to track users as they navigate the internet and data privacy concerns that arise because of this. Further research into this space has revealed advertisement blockers use the same technology to complete their task. New methods of completing this task should be developed.

This report describes how this system was designed and how the design will be implemented. The backend will utilise two main machine learning algorithms, word2vec and the inception model. The system will be used through a chrome plugin implemented in JavaScript. The middle layer will then connect these algorithms and front-end.

The final system will be a chrome plugin that states if there are advertisements on a website and presents a metric for how sure the system is.

Declaration

I hereby declare that the work described in this dissertation is, except where otherwise stated, entirely my own work and has not been submitted as an exercise for a degree at this or any other university.

Signed:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Student Name

Date

Acknowledgements

I would like to acknowledge the support I have received during the undertaking of this project.

My supervisor, Brian Gillespie, for guiding this project.

My family for their ability to translate my English into readable English.

And finally, my fellow students for the sense of comradery and support during my time at TUDublin.

Table of Contents

[1. Introduction 7](#_Toc35341514)

[1.1. Project Background 7](#_Toc35341515)

[1.2. Project Description 7](#_Toc35341516)

[1.3. Project Aims and Objectives 7](#_Toc35341517)

[1.4. Project Scope 8](#_Toc35341518)

[1.5. Thesis Roadmap 8](#_Toc35341519)

[2. Literature Review 9](#_Toc35341520)

[2.1. Introduction 9](#_Toc35341521)

[2.2. Alternative Existing Solutions to Your Problem 9](#_Toc35341522)

[2.3. Technologies you’ve researched 11](#_Toc35341523)

[2.4. Other Research you’ve done 14](#_Toc35341524)

[2.5. Existing Final Year Projects 15](#_Toc35341525)

[2.6. Conclusions 16](#_Toc35341526)

[3. Experiment Design 18](#_Toc35341527)

[3.1 Introduction 18](#_Toc35341528)

[3.2. Software Methodology 18](#_Toc35341529)

[3.3. Overview of System 18](#_Toc35341530)

[3.4. Front-End 18](#_Toc35341531)

[3.5. Middle-Tier 18](#_Toc35341532)

[3.6. Back-End 18](#_Toc35341533)

[3.7. Conclusions 18](#_Toc35341534)

[4. Experiment Development 19](#_Toc35341535)

[4.1. Introduction 19](#_Toc35341536)

[4.2. Software Development 19](#_Toc35341537)

[4.3. Front-End 19](#_Toc35341538)

[4.4. Middle-Tier 19](#_Toc35341539)

[4.5. Back-End 19](#_Toc35341540)

[4.6. Conclusions 19](#_Toc35341541)

[5. Testing and Evaluation 20](#_Toc35341542)

[5.1. Introduction 20](#_Toc35341543)

[5.2. System Testing 20](#_Toc35341544)

[5.3. System Evaluation 20](#_Toc35341545)

[5.4. Conclusions 20](#_Toc35341546)

[6. Conclusions and Future Work 21](#_Toc35341547)

[6.1. Introduction 21](#_Toc35341548)

[6.2. Conclusions 21](#_Toc35341549)

[6.3. Future Work 21](#_Toc35341550)

[Bibliography 22](#_Toc35341551)

# 1. Introduction

## Project Background

In January of 2019, Google proposed a change in their extension manifest version (v3) (1). The point of this update is to make extensions safer by adding greater permission controls. This however had the knock-on effect of “killing” ad block support according to Raymond Hill, the maintainer of uBlock origin and uMatrix (2). A conclusion therefore can be made that extensions cannot be used reliably to protect privacy while browser creators have a vested interest in creating targeted ads.

On the 25th of May 2018, the EUs General Data Protection Regulation (GDPR)(3) came into effect. GDPR has made it mandatory for websites to ask for consent to use the user’s data for personalizing content and ads. The problem here is often if you do not agree, you are unable to access the content on the website. In other countries the user is not given the option and consent is implied by the user using the service. This makes ad blocking necessary to protect the privacy of users who want to use such websites.

Observations were made of the turmoil between advertisement suppliers and advertisement blockers, as a result research to find other solutions to the Ad blocking problem commenced. Pi-hole, a DNS sinkhole for advertisements appeared as a promising solution. This program is simple to install but requires an always on server. The problem with this solution is it required prior knowledge of computing whereas browser plugins are a 1-click install meaning they are far simpler for the greater public to use.

## Project Description

To solve the problems mentioned above, a solution was proposed where the system uses a convolutional neural network model to Identify images in a website. The output will then be compared against what the text on the page is about. To compare the text and image classification output, word2vec a word embedding 2-layer neural network will be used decide what the text is about. If there is a miss match, we can assume that the suspect image is an advertisement. As time goes on advertisement providers are getting craftier or have the ability to remove the impact of the ad blocking software, as a result more ads are slipping through the cracks. This system would get around these issues by not having static ad lists that must be maintained.

Do a diagram

## Project Aims and Objectives

The goal of this project is to decide if a website has adverts or not. To achieve this objective the images and text on the page will be classified. The outcome of these two machine learning algorithms will then be compared against each other. This will be done by grouping these words based on what they are. For example, a dog is in the canine category were as classes will be in an item category. These are two separate groups, so it is likely the image is an advertisement. In this project, the concept will be refined and tested as a valid method of detection.

To achieve this a number of objectives must be completed. Step one is to gather relevant technology’s and test their suitability for the systems application.

## Project Scope

This project is about ad detection. Ad removal is out of scope for this project, the system will inform the user on how likely an image is or is not an advertisement. For the purpose of this system, the scope will be narrowed to the type of advertisements and text the system will deal with. This can be expanded depending on how accurate the system is at predicting the likely hood of the image being an advertisement. Starting with a small number of advertisement types will allow testing and validate if this is a relevant method of checking for advertisements.

As an application the system must be lightweight and fast. The interface should follow the concepts of universal design.

## Thesis Roadmap

One sentence summary of the following chapters This section will provide a summary of each of the chapters covered in this report.

*Research*

This chapter delves into existing methods of ad detection. Following this, an examination into current/existing technologies and data that will enable the completion of this project. Finally, this chapter will discuss any other research done for this project.

*Design*

This chapter explains the design and overall layout of the system. An explanation each layer of the system will be explained. This will include UML diagrams.

*Development*

The development chapter goes through the entire development process of the prototype system. The middle-tier and backend systems that where discussed in the Design and Research chapters are implemented here.

*Testing and Evaluation*

In this chapter I discuss the testing methodologies used and define the parameters for success.

*Redevelopment*

This chapter goes onto explain what steps must be taken to incorporate the technologies used in the prototype into the final system.

*Conclusions and Future Work*

This chapter reflects on the completed prototype system and research completed. The conclusions drawn from this work will be used to propel the development of this system.

# 2. Literature Review

## 2.1. Introduction

In this chapter, key technologies to the success of the system will be discussed. A Comparison to alternative existing solutions to this problem will be made.

## 2.2. Alternative Existing Solutions to Your Problem

*uBlock Origin*

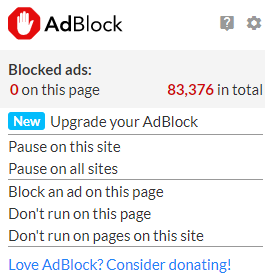
An efficient blocker: easy on memory and CPU footprint, and yet can load and enforce thousands more filters than other popular blockers out there (4). This is a complex interface for an app of this type. This is because the domains of each ad that are blocked are displayed. Control is limited to an on and off button. The app also has html element removing functionality.



*Figure 1- uBlock origin interface*

*AdBlock*

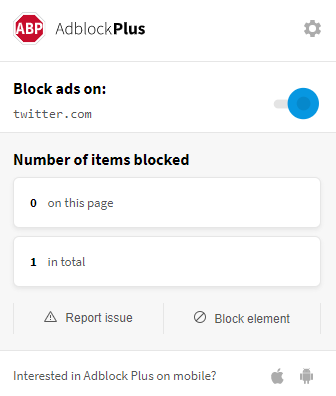
AdBlock (5) does exactly what it says on the tin. It supports the mainstream web browsers as well as android and iOS. The user interface displays the least information but allows the users to pause block and control running in one simple page.



*Figure 2- AdBlock interface*

*AdBlock Plus*

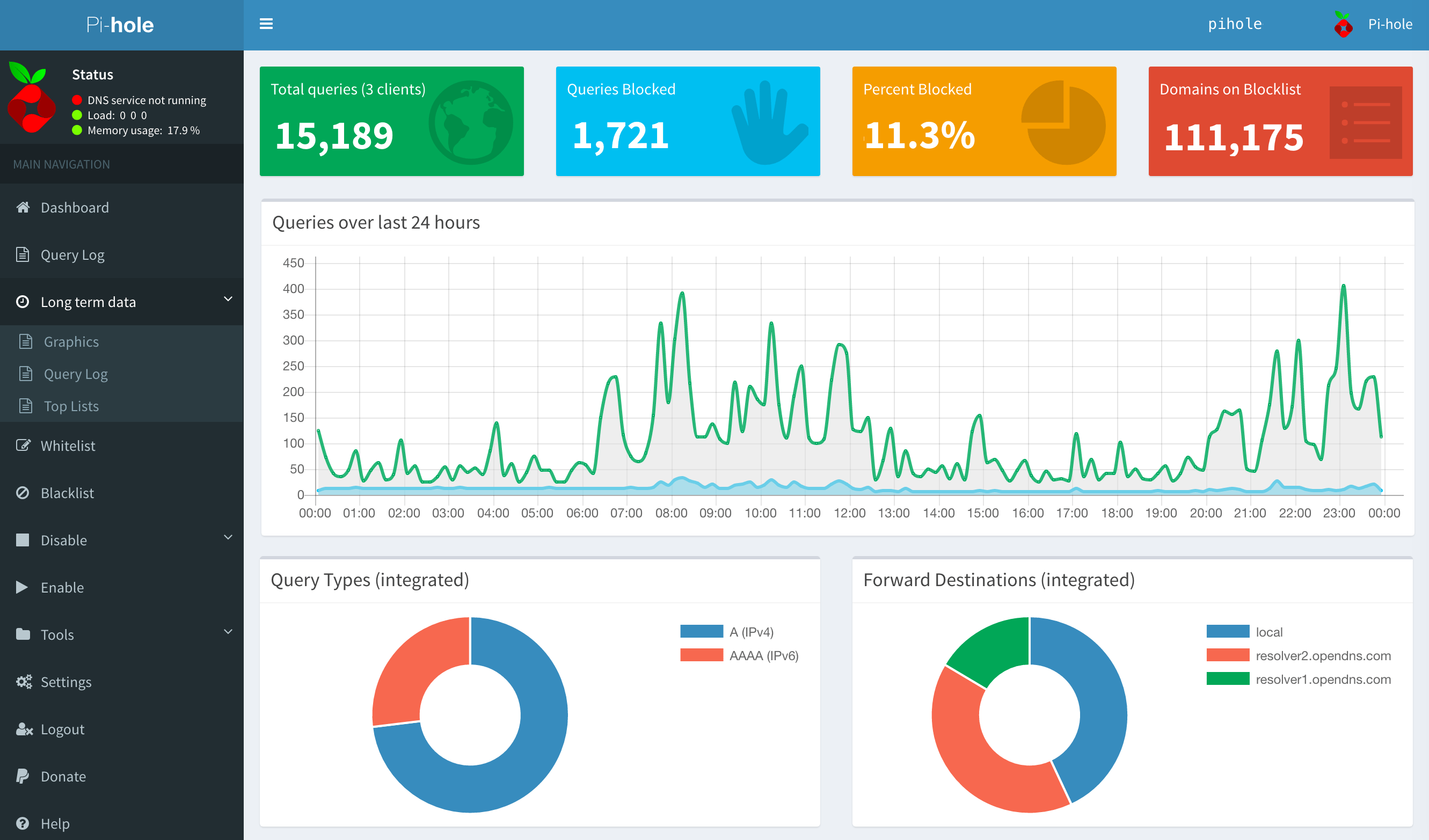
AdBlock plus (6) provides a distraction free browsing experience in the same was as others in this list. AdBlock plus is open-source and gives users access to privacy controls. The interface is very simple, this is a toggle for blocking ads on the current website and a display for advertisements blocked.



*Figure 3- AdBlock Plus*

*Pi-hole*

Pi-hole (7) is a DNS sinkhole that is used to provide network wide adblocking. This is another system that uses static lists to find and block advertisements. The system is harder to setup then its counterparts as it is hosted on an internal server and client DNS setting must be changed to point to the internal server. The system can also double as a DHCP server.



*Figure 4- Pi-hole dashboard*

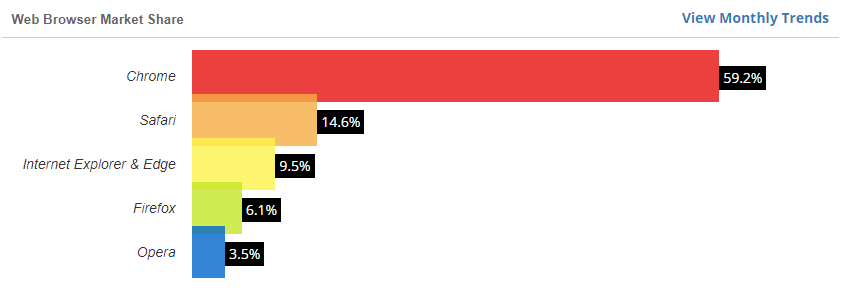
*Conclusion*

These programs all use a variation of static lists to find adverts. The most common are plugins for common browsers i.e. Firefox, chrome, etc. All these solutions have very simple interfaces that allow the user to see the total number of advertisements blocked on the site. They also allow the user to enable and disable on the fly. All the plugin mentioned have 10,000,000+ users on google chrome making this a very sought-after functionality.

## 2.3. Technologies you’ve researched

The following are possible platforms:

*Chrome*

Chrome is a popular web browser (8). Chrome currently has about 60 percent of the browser market according to statcounter and w3counter. These statistics are collected from all websites that use W3Counters or statcounter, these websites collect telemetry using JavaScript allowing insights into the website’s performance. Google are a major supplier of advertisements, so have a vested interest in crippling the current systems (9). 

*Figure 5- W3counter report*

*Firefox*

Firefox (10) does not have anywhere near the popularity of chrome. Firefox has a reputation for user privacy and in 2019 have rolled out features such as a VPN and have removed plugins that collect excessive user data. By default, 3rd party tracking cookies and crypto miners are blocked. Firefox is backed by a non-profit.

*Conclusion*

In conclusion, chrome is the market leader for web browsers. A method that does not uses static lists will be most useful on this platform as this is where the advertisement skirmish is happening. Firefox actively supports user privacy meaning that if the system is successful it should be ported to this platform in the future.

The following are possible programming languages:

*Java script*

This is the go-to language for extension development along with html and CSS. Java script can be used to modify the DOM on websites this will be used to highlight or completely remove the ads.

*Python*

Python is a high-level interpreted language that emphasised code reliability. Python supports many open source frameworks, library’s and development tools. Python also allows for rapid prototyping due to the small size of the code. Applications like conda or virtualenv allow for python programs to be isolated from other projects and help manage their dependencies.

*C++*

C++ is a general-purpose compiled language. C++ is used mainly in systems programming and embed systems because of its features and security. C++ has support for library’s and frameworks.

*Conclusion*

In conclusion, python’s ability to produce rapid prototypes and mature library’s make it a good language for this system. Python is often used in machine learning and web development tasks.

The following are possible library’s and models:

*Word2vec*

This is a method of converting words to a machine-readable format. Normally to deal with categorical features like words one must embed them into a numerical format but in doing so it loses how the words relate to each other. Word2vec preserves the relationship between the words. Word2vec comes in two flavours, the skip-gram method of word2vec weighs nearby context words more heavily than more distant context words and Bag of words which the order of context words does not influence prediction. (11)

*Inception*

Inception is a convolutional network that classifies images. Unlike its competitors it goes wider instead of deeper by having multiple convolutions on the same level. Using a pre trained model as opposed to building a new model makes sense when trying to solve this problem as these models are optimized and trained well past anything one person can manage on their own. (12)

*WordNet*

WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. (13) The linking of the items in the wordnet database and our ability to see what Synset the word is part of will allow the system to make meaningful comparisons between the output of the convolutional neural network and text analysis systems

TensorFlow Lite

This allows TensorFlow models on mobile, embedded, and IoT devices (14). Lite does this by using pre-trained models. These models are designed to execute efficiently on mobile and other embedded devices with limited compute and memory resources. TensorFlow Lite has a python implementation this should allow me to use python as the backend of my extension/website. There is also a JavaScript implementation, this could be used to run the image r

The following are possible web frameworks:

*Flask and flask restful*

Flask (15) is a micro web framework for python that provide the means to create a web application. It is free and open source. Flask restful builds on this functionality allowing developers to create REST APIs. A rest API may be necessary for this system depending on what portions of this project can be ran in browser. Flask is a micro framework meaning that functionality like logging in and Restful APIs are added in separate libraries or must be implemented by the developer.

#### Django

Django is a web framework for python that includes all the functionality needed to scale and secure a website. Django like flask is free and open source but is more feature rich out of the box.

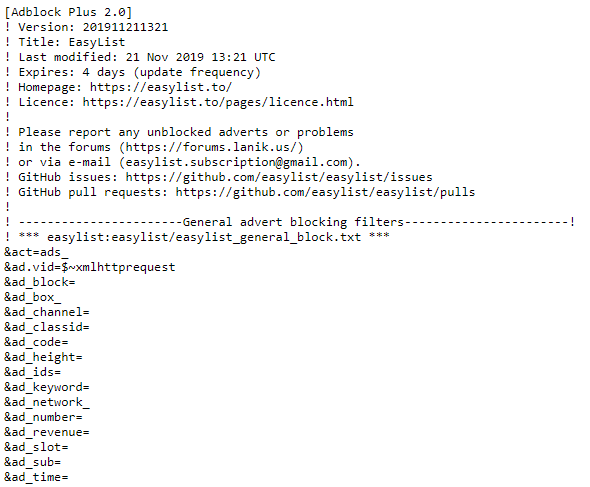
*Conclusion*

Flask is a far simpler and more flexible as it sets out to be a mini web framework and nothing more. Whereas Django aims to be an all-inclusive platform. As of now I only require a method of communicating between the middle layer and backend because of this the lighter simpler platform will be better for this project.

## 2.4. Other Research you’ve done

*Ad lists*

These are lists made and maintained by the AdBlock community. A popular Ad list, EasyList (16) comes pre-packed and/or tested with many adblockers (AdBlock Plus, AdBlock and uBlock Origin). EasyList is a primary filter list for removing most adverts from international webpages, including unwanted frames, images and objects. It is the basis for many other filter lists.



#### Draw.io - Diagrams.net

Draw.io is an online diagram editor. The application is browser-based and free to use. This application will be utilised for medium fidelity prototypes of the UI of the system. Another application is diagrams for the design and ideas of some systems. As such draw.io will be utilised in the explanation of this system.

## 2.5. Existing Final Year Projects

Here are some final year projects from previous years that are relevant to this project. These projects are both machine learning projects and as such share some similarities to this project in their development style, technologies and methods used.

*Title:* Euro Coin Classification. Using Image Processing & Machine Learning

*Student:* Yumin Chen

*Description (brief):*

This project aims to investigate the visual features of euro coins and develop models that can represent different natural images. The objective of this project is to investigate the use of machine learning specifically in the context of euro coin classification.

*What is complex in this project*

Classifying the coin is the difficult part of the project. This is done using both image processing and machine learning. The student implements different algorithms and evaluates them.

*What technical architecture was used*

OpenCV, Cordova, machine learning and data mining techniques.

*Explain key strengths and weaknesses of this project, as you see it.*

The student tested a multitude of ml algorithms and evaluated them against each other. I see this as a strength. The student uses Rapid Application Development methodology, prototypes were made quickly but the problem of immediately starting to code is an issue. This weakness in the development style, extra care should be taken in the planning stage for this sort of project.

*Title:* Detecting Bot Twitter Accounts using Machine Learning

*Student:* Emmet Hanratty

*Description (brief):*

In this project the student creates a system to detect and evaluate whether a user account is a bot. The system manages good accuracy, and this will be increased as the data set expands. They system also creates graphs on the data used in prediction.

*What is complex in this project*

Training a sentiment model based on tweet data. This decides if a tweet is positive, natural or negative.

*What technical architecture was used*

Python using the Django framework, MySQL, Twitter API and Git

*Explain key strengths and weaknesses of this project, as you see it*.

The project is not multi-threaded meaning it is slow in the pre-processing stages. Twitter have a limit on the amount of calls you can make to their API, meaning after a certain number of requests the system is unusable for the rest of the day.

The system has good accuracy when predicting the state of an account and this accuracy will grow over time.

2.6. Conclusions

From looking at the current popular solutions for advertisement blocking we can see they all employ a similar approach for this task. With certain companies having a vested interest in curtailing this technology for their own profit and said companies having the means of doing so other solutions should be developed. To develop a solution to this problem possible technologies were researched, and existing Final year projects were looked at to provide guidance.

# 3. Experiment Design

## 3.1 Introduction

In this chapter,

## 3.2. Software Methodology

*Waterfall Methodology*

The waterfall model (17) is a phase by phase method of developing software. The current phase must be completed before commencing the next phase. Once a phase is finished it cannot be revisited till the end of the project.



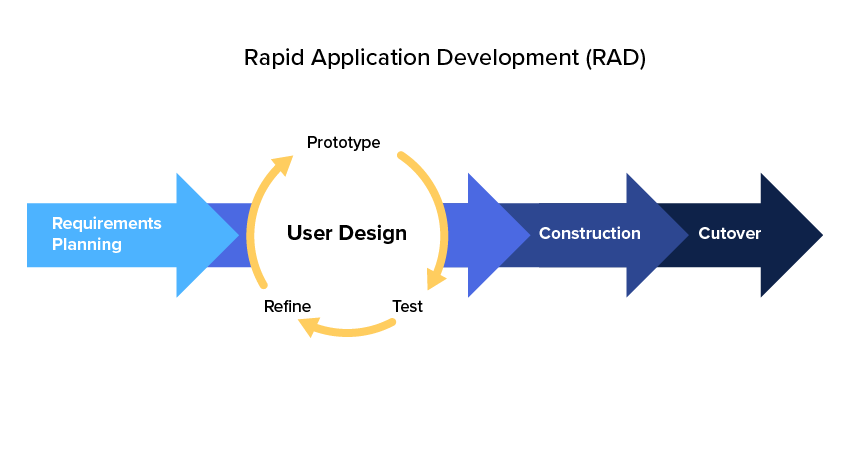
*Figure 7- Waterfall model*

An advantage of this model includes clear and understandable milestones. This makes the stages of the process and advancements very clear.

A disadvantage of the waterfall model is that is difficult to redevelop and redesign when the goalpost changes. Once started you are locked into this mode of development till the maintenance phase. On completion of the project, the application may or may not be suitable for its purpose.

*Rapid application development*

Rapid application development (18) is a form of agile software development. This methodology will allow rapid testing and creation of new approaches.



*Figure 8- Rapid Application Development*

The main advantage of the RAD methodology is that time between prototypes and their next iteration is short. This allows for rapid development over a shorter period.

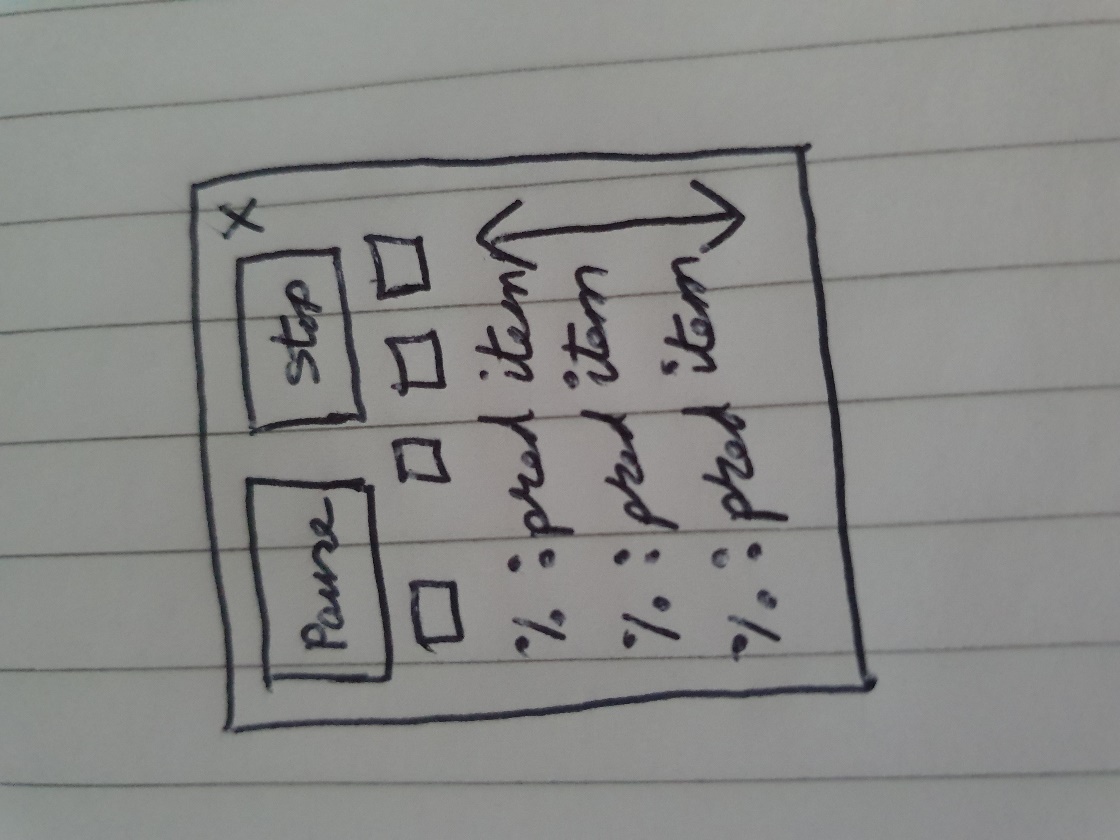
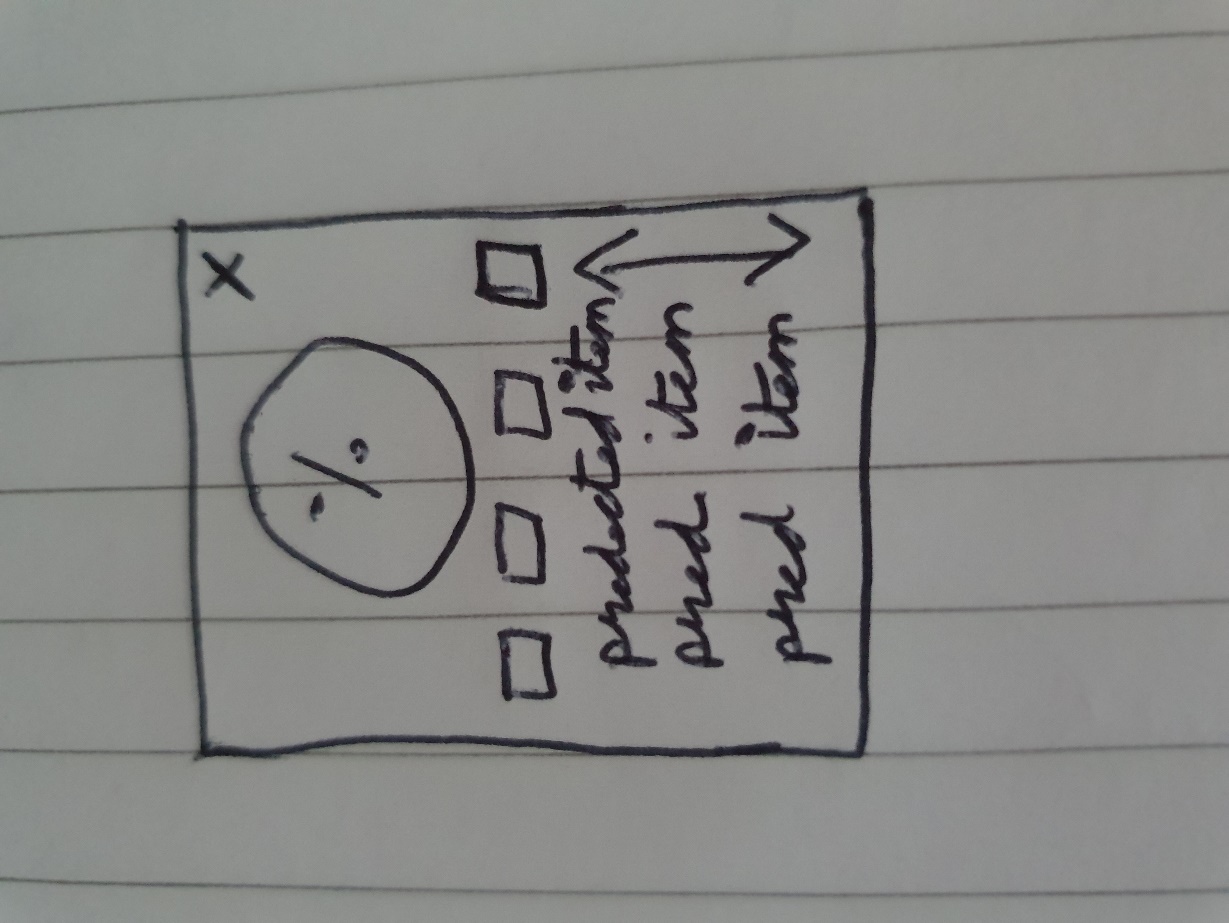
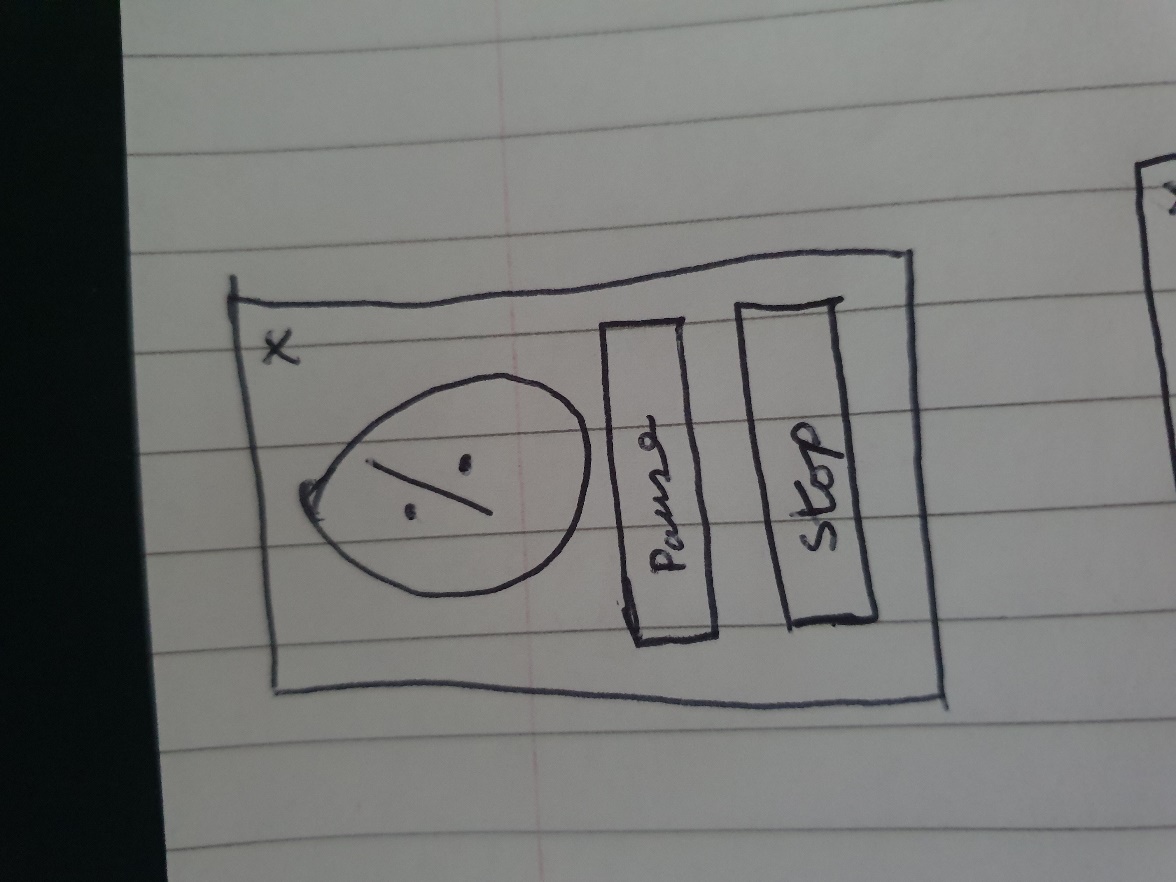
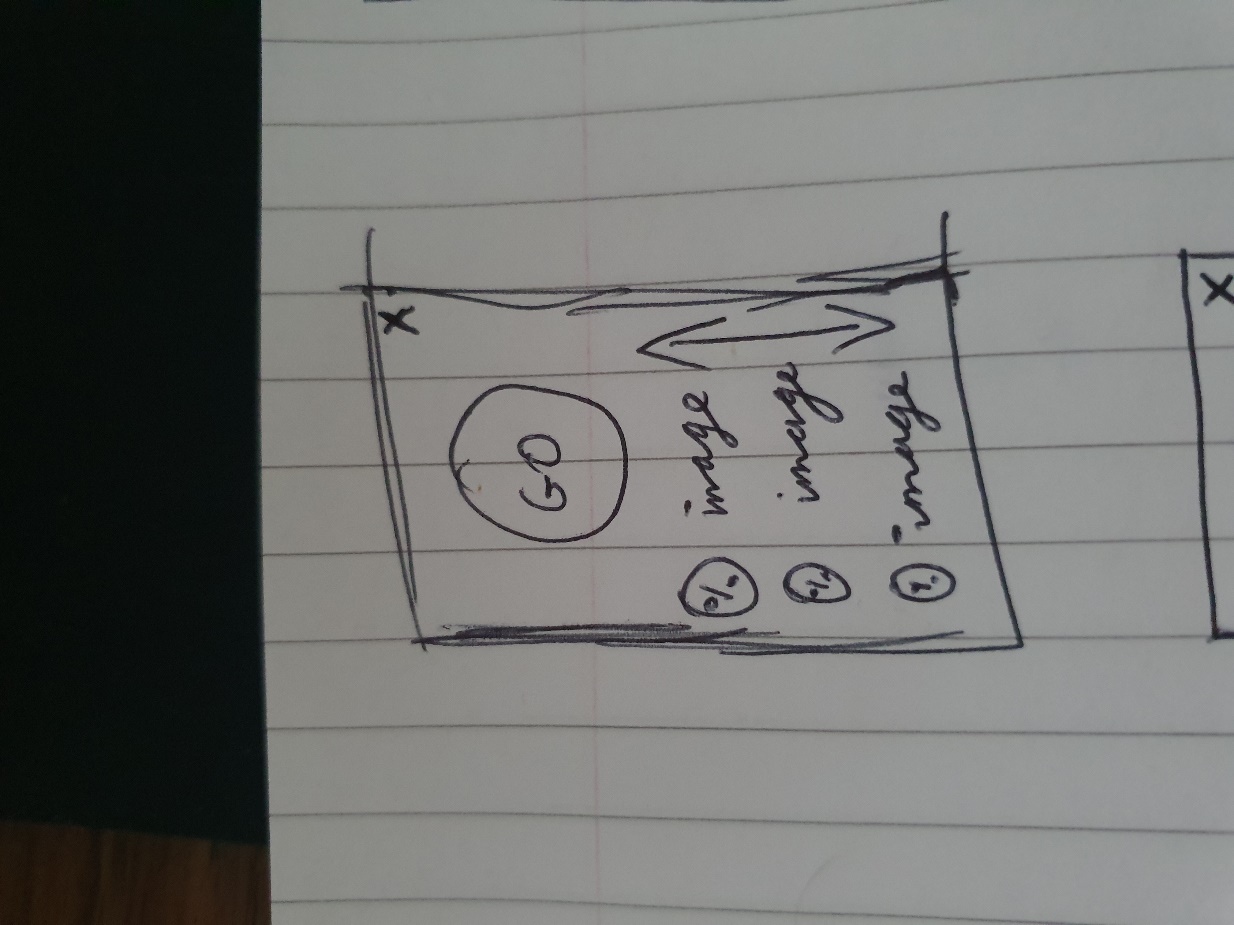
The disadvantages of RAD are mainly based around applying the methodology to larger teams. As this is a solo project this is not a problem. Another disadvantage of this methodology is it is not suitable for larger projects.

## 3.3. Overview of System

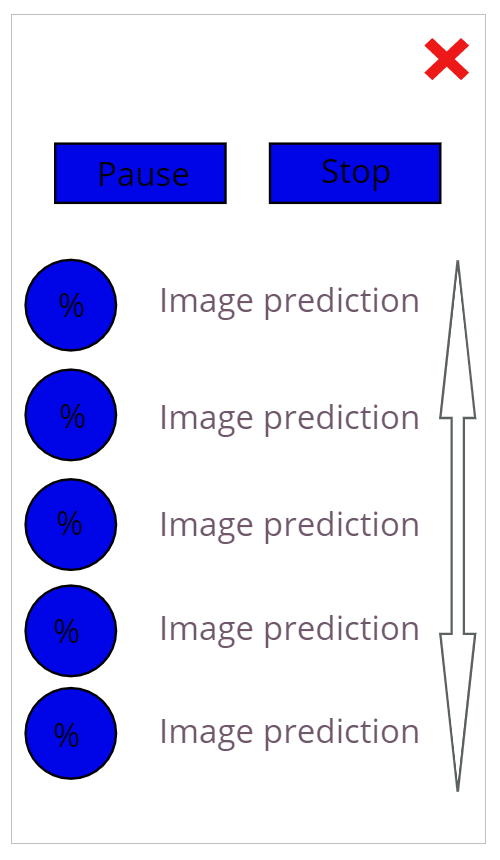
Using the Rapid application development method of software design the first step is define what the system aims to accomplish, this has been completed in part 1.3 of this report. The second stage of RAD is to prototype designs and develop these prototypes. The general approach to the prototype phase was to divide the system into manageable pieces. These pieces were then built, refined and tested. After Unit testing which will serve as the feedback, these pieces will be grouped together to create the final system.

## 3.4. Front-End

The system will have a similar front end to its alternative solutions. This will be written in html, CSS and JavaScript as these are the default languages used to write extensions for chrome. Across these solutions, the number of ads blocked, a way to stop and restart the application are universal. Below are possible designs for the interface of the application.



The next step was to create medium fidelity prototypes. The front end will we simple with one to two different pages. The prototype displays the functionality of the interface and its overall layout. A simple easy to understand interface will server this system the best as it will need to be accessible to all types of users.



The final design will.

The front end is also responsible for the parsing and gathering of the data. This is then posted to the Flask API.

## 3.5. Middle-Tier

This is the flask API responsible for interacting with the backend systems and the front end.

The API will utilise both get and post requests. The get requests will be responsible for explaining to the user what and how the API should be used. Posting to the API on different routes will allow the chrome extension to interact with the backend systems. This separation will allow lower spec computers to take advantage of these algorithms.

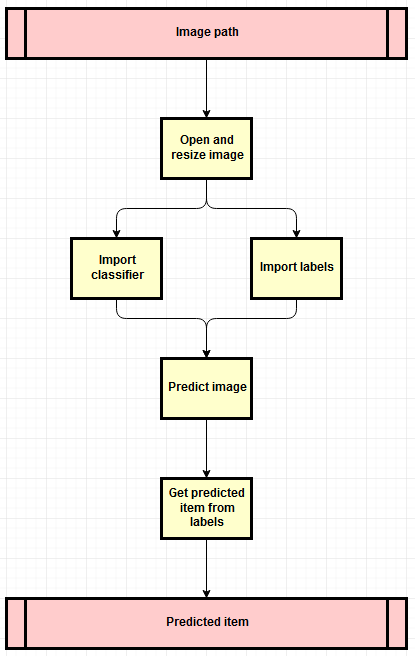
## 3.6. Back-End

The backed will consist of 3 systems, image classification, word embeddings and .

*Image classifier*

The image classification system will be based on the inception model. This model is already trained on ImageNet making its predictions already quite accurate. The plan to expand on this model is to retrain it with logos of common companies making it more suitable for these purposes. This backend system may also be moved to

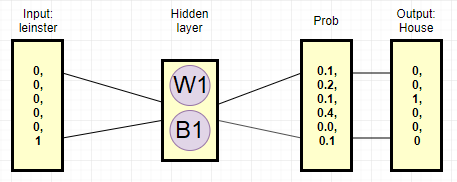
Below is a possible structure for the image prediction.



*Figure 12 - Image Classification Design*

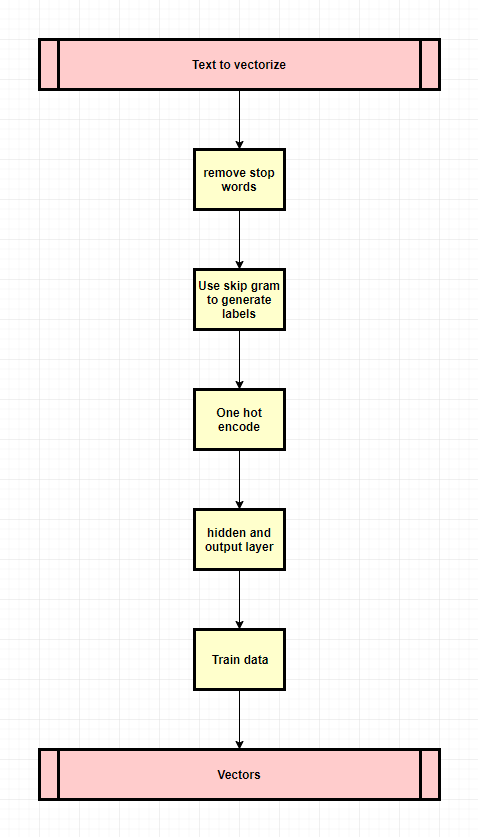
#### Word2vec

The word2vec (19) portion of the back end will run on a server as it is to intensive to run on the clients. Each new document will have to be trained separately as each document will produce a different set of vectors. Word2vec outputs word embeddings by predicting the neighbouring words given a word. Word2vec is a 3-layer neural network which consists of an input layer, a hidden layer and an output layer.



*Figure 13- Word2vec structure*

Training the model to predict the next word will never give high accuracy at the task. But the hidden layer which is the vectors used to predict the outcome give us the relationship of the words to each other.



*Figure 14 - word2vec Design*

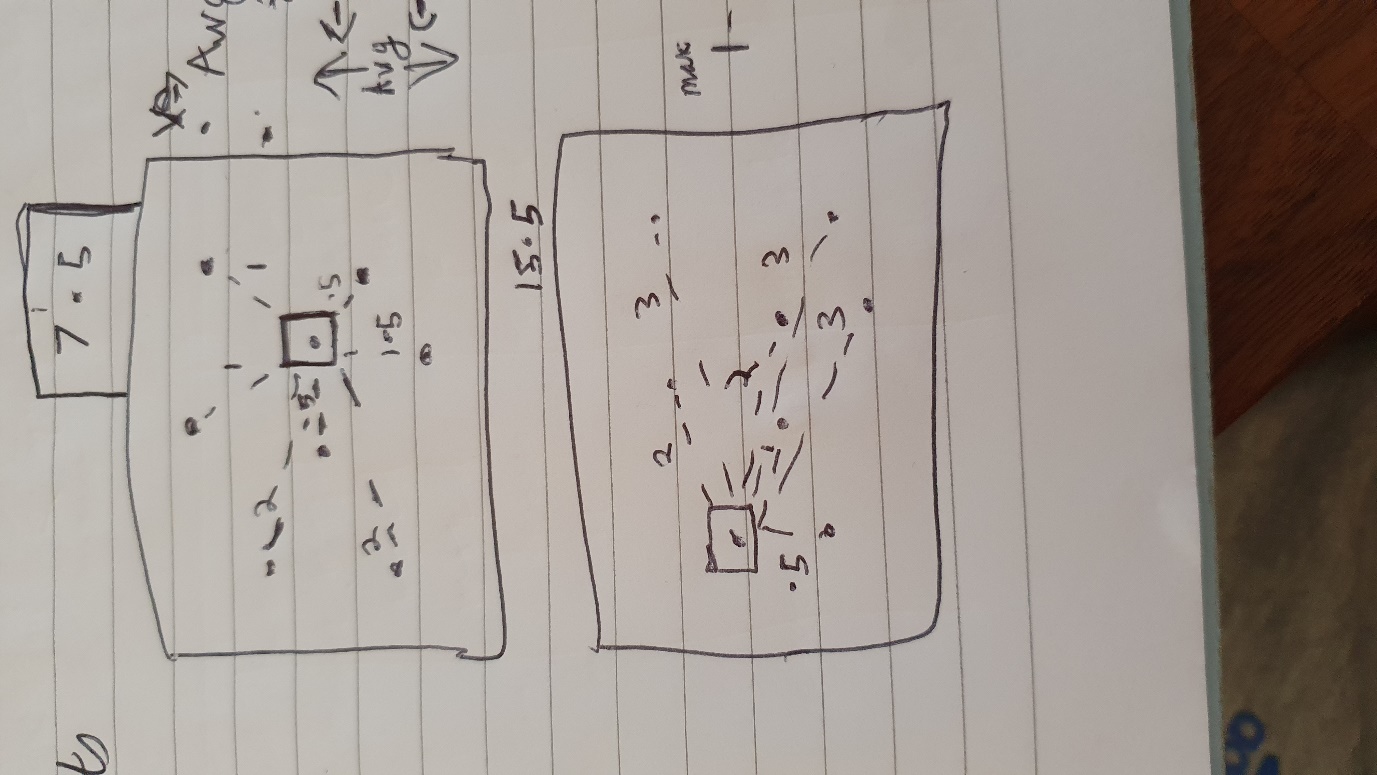
#### Compare

This portion of the backend will be responsible for comparing the outputs of the image classification and word2vec. This logic will come in two steps, first the system will decide if the image could be an advertisement or not. This will be a Boolean value based of if the word is in the text. As words can be different but mean the same thing, querying the word net database is necessary. This will allow the system to match a greater variety of words. To do this synsets and lemmas are utilized. Synsets are sets of synonyms and lemmas are the canonical form of the word. The primary method of discerning if words are part of the same category will be the synsets but if this fails lemmas will be used to substitute the data. Below is an example of a synset relationship.

A close up of text on a white background

Description automatically generated

In this example both wolf and dog share the synset or category of canine. The system gets the values of 'canine', 'unpleasant\_woman', 'chap', 'villain', 'sausage', 'catch', 'support' and 'pursue' for dog. And the values of 'canine', 'womanizer', 'attacker' and 'eat' for wolf. Both lists have the values of canine so therefore the image and text wold have the same classification making the image being an advertisement unlikely. But as stated before this is a Boolean yes or no, dog could have been mentioned in a different context like hot dog. As such this compassion only serves as a starting point to decide how likely this is to be an advertisement.

The second stage of the compare logic is to decide how likely the possible advertisement is to be an actual advertisement. To do this the output of the word2vec algorithm is used. This is a two-dimensional graph of how each word in the text relates to each other. Using the Euclidean distance, we can compare how connected a word is to the document. Therefore, a more connected word will have a smaller total value for the lengths of its neighbouring points. This is a necessary part of this system as the text of a web site may mention something that does not pertain to what the said text is about. For example, a text may mention that they have a new line of clothing and you should buy them. This text will not be as interconnected in the word2vec graph, giving it a higher value meaning it is less likely to be what the subject is about.

In the above diagram this idea is shown, the middle point is more connected meaning it will get a lower score whereas the second diagram the point is on the edge meaning the other points are further away making the value higher.

There are a few ways to calculate the distance between two points such as the Euclidean distance, Manhattan distance and Chebyshev Distance. Each of these methods have different quirks as they approach the problem differently. Because the Euclidean distance is the easiest to intuitively understand it was chosen as the method for calculating this metric, but this requires more mathematical operations so in future versions it may be quicker to use the Manhattan distance as this method requires less operations. The Euclidean distance can be equated to traveling along the hypotenuse of a right-angled triangle whereas the Manhattan distance can be equated to traveling along the other two sides.

## 3.7. Conclusions

In this chapter, the three main back end technologies were discussed was given for each of t, and an overall design hem. The middle tier will be responsible for connecting the backend to the frontend and translating the gathered information into meaningful data. The design of the UI was explored and fleshed out.

# 4. Experiment Development

## 4.1. Introduction

From my research and design, I have decided on a few technologies to use for this system. These technologies may change depending on how this experiment develops. Upon entering a new site scraping the DOM for the relevant data (the images and text of the site). This data is then shipped of to the flask API which is responsible for running the necessary backend functions

## 4.2. Software Development

The first step to developing any system is to set up version control. This project is being managed through GitHub. As the prototype of this system already has a Git repository, we will be extending from that. In order to keep different versions of python libraries from interfering with each other it is necessary to utilise conda or virtualenv. For this project virtualenv was utilized as it was pre-installed, and the developer has experience with this method of package management. Visual studio code has a robust and easy to use implantation of version control along with good python virtual environments implementation as such this is the chosen IDE.

## 4.3. Front-End

The front-end interface will be implemented in html, CSS and JavaScript. The main html file will be responsible for showing the user information about the system. Upon entering a new website, the chrome extension will run the script responsible for collecting and shipping the data to the middle tier Flask API. The returned information will then be displayed to the user. The chrome extension consists of a manifest file, the popup HTML and CSS and any scripts that should be run.

The manifest file is responsible for providing chrome with important information. The required information by chrome is the manifest version, name and version. The name is simply the name of the extension, the version is the version of the chrome extension and the manifest version is the version of the manifest file which as of Chrome 18 should be set to the integer 2.

## 4.4. Middle-Tier

The middle tier of this system is a flask API. This system is responsible for accepting the POST and GET requests from the chrome extension and passing this data to the backend scripts that classify the data and decide if said data constitutes an advertisement or not. Each of these methods have a different route in the flask API. These routes along with the URL state the type of http traffic that the method will accept.

The POST methods of this system import the process management methods of the classifier scripts. When a POST request is made to the URL, Flask will run the method associated with that route which will call the necessary backend script. This data is then returned to flask in a dictionary format and sent back to the chrome extension.

A screenshot of a cell phone

Description automatically generatedThe above code is responsible for handling the image prediction. To receive images from the client chrome extension, the images are encoded as base 64. To decode these coded strings into an image, the system must strip the identifying string from the coded string. The resulting string is then decoded from base 64. In order to get an image from this the system must convert it into a byte array. As the image declassified takes 3-layer images the image must be converted to the RGB format. The image is also resized to 224 by 224. The image is then saved for testing purposes. The image is then passed to the back-end image classifier and the result of the image classification is send to the client.

The GET methods are responsible for explaining what the corresponding POST method does and the way in which the data should be structured for that method to be used. The below code is responsible for responding to the get request. When the get request is made the api displays information on what the route is used for, how it should be used and what the expected return type will be.

A screenshot of a cell phone

Description automatically generated

<https://developer.mozilla.org/en-US/docs/Web/HTTP/CORS> As we are using Flask as a fetch API, CORS (Cross-Origin Resource Sharing) must be used for the chrome extension to communicate with said API. In order to do this, we must import the CORS flask library, set the CORS headers in our flask app to content type and before running the app convert into CORS. By default, flask uses multiple threads to respond to requests. This functionality enables the flask API to respond to requests simultaneously. For this experiment threading must be disabled, this makes flask queue the requests. The backend algorithms utilised use a large amount of ram as such running more than one algorithm at one time will crash the server. Disabling threading is a quick fix for this for the purposes of this experiment. The development server included in flask is not capable of managing threading and multiprocessing in order to get access to those features it is necessary to deploy a production A screen shot of a person

Description automatically generatedenvironment, this is out of scope for the current system.

## 4.5. Back-End

Word2vec, image classifier , wordnet.

#### Ram usage

The image classifier and word2vec use a large amount of ram. After these systems finishes, they do not release the ram. The research conducted and other options tried to solve this problem are explained in the testing chapter. The solution settled upon is to create a new process for these systems. As the flask development web server is a development server it does not have the functions necessary to manage processes. As such the pythonic way of creating and managing new processes is used. When creating new processes inside the running api the system will hang while it tries to spin up an infinite number of processes. To solve this issue the new processes must be created before the api is started or while the api is executing a function. Trying the former method, the processes can be created and utilised with the intended effect but when closing a process the system lacks the ability to create a new process in its place, leading to the system running out of new processes after n number of calls to the api. As such the latter method of utilising threads was used in the final system. To utilise processes in python the multiprocessing library is utilised. Creating a new process is simple as all that is needed is the process method. This method takes a target function which will be called in the new process and arguments that will be passed to the new process. In order to get data back from the process, the manager class is utilised to get this data back. The manager class allows the creation of dictionary proxies. The dictionary proxy is then utilised to get data back from the process. This allows the contents of the dictionary to be propagated through all the processed with access to the dictionary. As the proxy dictionary is not a dictionary, to return data as a dictionary the system takes the first value of the proxy dictionary which is a python dictionary. The next step is to start the child process and join waits for this child process to terminate. The contents of the dictionary are then returned to the flask api. Because this process is necessary for two of the back-end systems, word2vec and image classification, this explained here. Below is the code used to complete this task in image classification.

A screenshot of a cell phone

Description automatically generated

#### image classifier

#### Word2vec

In order to determine the likely hood of a word being an advertisement we must decide on a method to quantify what the text is about. The chosen method for this experiment, as disused in the design chapter, is word2vec as this allows us to change the document into vectors on a graph. This graph shows us each of the words in the text in relation to each of the other words. The skip gram method was chosen as this method predicts the context rather than the word.

A screen shot of a computer

Description automatically generatedThe first steps to implementing this algorithm is cleaning up the data. To do this the system uses a regular expression to find all non-alphanumeric characters and removes them. In order to preserve the structure of the sentences in the text, the text will be split into a list by “. Space” before the regular expression is used to clear up the non-alphanumeric characters. In this portion the trailing new line characters and capital letters are also handled.

The next step is to remove the stop words from the text. Stop words are a set of commonly used words in the English language. The reason why these words should be removed is as the removal of these words allows the system to focus on more important words to the text. For the purposes of this experiment the system uses a limited number of stop words though there are many more words that can be added. Removing more stop words may have a negative or positive effect on the overall system as such the system utilises a small number of stop words. Below is the code utilised to accomplish the removal of the stop words.

A screenshot of a cell phone

Description automatically generated

Next the system creates a list of words in the text. To do this the set function is utilized removing the duplicated words. This list in then utilized to create a mapping of a word to an integer. To create this dictionary the enumerate function is utilised. This function returns the location of the item and the item and these items are then inserted into the dictionary.

The next function creates the training data used for the word2vec algorithm using the cleaned and stop word less data and the enumerated words. The structure of the training data is a list of word pairs. The word pairs are generated by scanning through a sentence and each word index plus and minus 1 value till the window size is hit. Below is an example of some generated data. As you can see the stop words are removed by earlier processes as such “handful” is the first item in the list. “handful” is related to “of” and “major”. This is repeated for the entire list of sentences.

A screenshot of a cell phone

Description automatically generated

To generate this data, it is necessary to enumerate the words list. This gives the system the index of the words of the list. Using this index, a range of words are selected and if the words selected are not the current word this relationship is added to the training data. The below code is responsible for executing the above logic.

A screenshot of a cell phone

Description automatically generated

The next step is to convert the data into one hot encoded data.

To do this the system gets the count of the words. This is used to set the size of the encoded data which are all zeros. The integer value of the word that was generated earlier in the code is then used to set that index to 1. This is done for all words, giving each its own encoded value. As in the below code.

A screen shot of a social media post

Description automatically generated

The next step is to generate the computational graph which is done through TensorFlow.

#### Compare

Comparing the outputs of the afore mentioned algorithms is done using the wordnet database and by using Euclidean metrics. As discussed in the design chapter it is important to be able to compare the outputs of these systems.

In order to query the wordnet database the word must be cleaned of any non-alphanumeric values. This is done using the re library and the following regular expression '([^\s\w]|\_)+'. This expression is compiled so it can be used against the submitted word. The word is then queried against the wordnet database using the synsets method. The returned object is a list of the synsets of that word. This is then looped through and the hypernyms method is used to query the synset for the category of the word. The category is then returned in the wordnet format this is then stripped of said format. To do this the split method is utilised, using the first full stop as the item to upon. Taking the first item from the split, using the lower method to lowercase the word and removing the first nine letters from the final output as wordnet prepends “hypernym” to the beginning of the output word. These hypernyms are the added to the output list. This logic is executed in code form below.

A screenshot of a cell phone

Description automatically generated

The next step in the method is to have a backup if this fails, as discussed in the design chapter if there are no synsets the system falls back on lemmas which are the dictionary form of the word. then the system filters out any empty, null value and original words. The returned value is a list of similar or synset words to the input word. This is completed by the below code.

A screenshot of a computer screen

Description automatically generated

The next step is to deal with passing double words to the system. As discussed in the design and testing chapters this can happen as the labels of the image processing system can be double words like red wine. As red wine can be used to represent the colour or alcohol in speech it is necessary to utilise both words in the matching process. To do this the system splits on spaces, calls the wordnet database querying function and concatenates the output of the wordnet database.

A screenshot of a cell phone

Description automatically generated

To handle a list of words at once the convert to wordnet function returns a list of the converted words in a list. To get these words a lambda function is created that passes words into the double word handling function. The map method is used to map this lambda function to input words. Below is the necessary code to perform this operation.

A picture containing black

Description automatically generated

To check if a list of image labels contains an advertisement the function discussed above is called. The output of this function is then flattened to a list and the original words are re added to the list. If the image list is loop through and if the image is in the c words list the value of true or false is appended to a new list. The outputted list then can be counted for the amount of false values and if there are no false values that means there are no advertisements and if there are advertisements true will be returned.

A screen shot of a smart phone

Description automatically generated

Now that the system knows that there could be advertisements it must work out how likely this is. To do this the vectors from the word2vec system is used. The first step to this is to implement the Euclidean distance method as a function. This function will take in four values being the two-dimensional points and will return the distance between these two points. As stated earlier in the design chapter Euclidean distance was chosen for its simplicity to understand nut this method required more mathematical operations than other methods. The Euclidean distance is simply x2 – x1 squared plus y2 – y1 squared and then this value is square rooted. Below is the code to complete the Euclidean distance method.

A close up of a logo

Description automatically generated

## 4.6. Conclusions

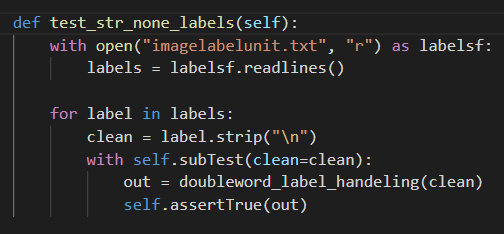
# 5. Testing and Evaluation

## 5.1. Introduction

In this chapter, the explanation of how unit testing occurred and how the final system was evaluated.

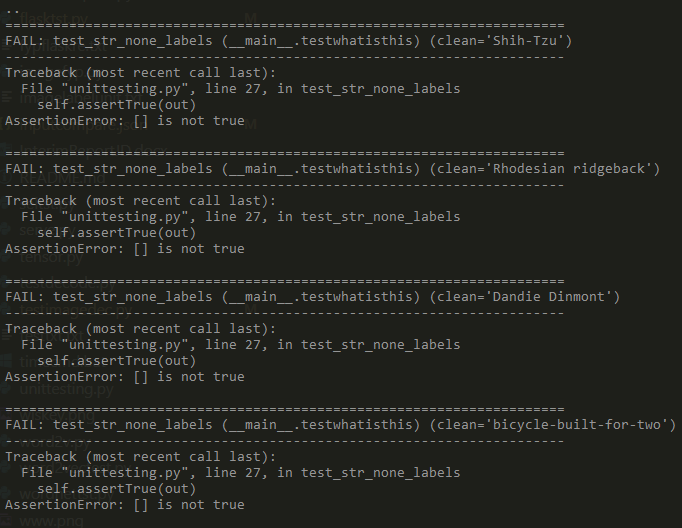
## 5.2. System Testing

In software development, testing is an important part of the development process. In the RAD methodology this is especially so. All software has bugs, to mitigate this testing after iterating is essential. Through out the development of this project both unit testing and ad hoc testing to validate the iterations of the programme.

Unit testing was most effective when applied to the code that decides what a word is. Using the labels of the image classification algorithm this function was tested and various changes were made because of this testing. For example, the original function did not consider double words. This was never a consideration as the word2vec model ensures a one-word format whereas in the image classifier 400 words are double words like red wine. The bug would not have been detected till much later or at all if not for the unit testing.

The above code is responsible for one of the subtests. The imagelabelunit.txt contains the labels of the image classifier, this is read in and cleaned up. For every word in this list, the wordnet database will be queried if there is no output then the query failed meaning there is a problem. The failure will be reported in the output of the unit test script.

In order to execute the above code, the unit test framework is utilized. The testing functions are incapsulated in a class. This class has the unittest.TestCase passed to it which sub classes unittest.TestCase. Contained in this new class is the methods for testing the system. These methods assert something to be True, False, Equal, etc. This creates the test cases that will be run. In order to run the test script, the main() method is called. This method is used to create a command line interface for the script as seen in the below image. The first 2 dots mean succeeded tests whereas a Fail is a test fail. The failures show what test class failed, what method failed, and the variables used. From this information, the reasons for the failure can be mitigated.



As seen above, the current state of the unit test is nine sub checks fail. Many of the words that fail are not in the wordnet database or are irrelevant for the evaluation of this experiment as such this is unnecessary to solve this problem.

An ad hoc method of testing was required for the Flask API as with this application it either works or it doesn’t. To do this testing the curl tool was utilised as the client plugin had not been developed at the time. This ended up being a poor method of testing as the chrome plugin utilises CORS which is used when a web app must query a fetch API. <https://developer.mozilla.org/en-US/docs/Web/HTTP/CORS> The apparent successful testing of the API meant that when developing the POST functionally of the chrome extension the problems experienced seemed to purely be on the extension side of the application. After implementing the CORS package in the flask API and adding CORS mode, accept header and body json this issue was solved.

Another problem in testing the API in this manner is was in running the algorithms one at a time and then closing the process. This meant that when running the scripts in the flask API the process wold never close witch is the intended functionality. But the methods to free the resources are not guaranteed to function. These methods include TensorFlow’s build in methods like reset\_default\_graph() and close() but also using library’s like numba which have a Cuda implementation to select the GPU and close the current thread. As such, the best way to manage the memory consumption of these scripts is to create a new process and when they are finished processing the data end said process freeing the memory back to the system. This is guaranteed to happen as the process had ended. <http://numba.pydata.org/numba-doc/0.13/CUDADevice.html>

## 5.3. System Evaluation

5 chosen Websites and display results.

## 5.4. Conclusions

# 6. Conclusions and Future Work

## 6.1. Introduction

## 6.2. Conclusions

## 6.3. Future Work

#### Browser image classification

#### Retrain the inception model

The default model is the first runner up for the ImageNet Large Scale Visual Recognition Challenge. ImageNet is an image database organized according to the WordNet hierarchy. Retraining the model for more advertisement-based images would improve the accuracy for this application of the classifier. These advertisement-based images could be company logos. The logo may not be the main thing the image classifier identifies but if the prediction has a high enough percentage the system can be sure that the image is an advertisement. The image can then be removed without having to compare it with the text. This will make the system faster and more efficient as the main system will not have to be queried if the image classification is done in browser.

#### Lan implementation

The current implementation of this project is designed to run on a webserver and to be queried by a chrome plugin this idea came from current implementations of AdBlock in browsers. Other methods to solve this problem exist like Pi-hole as mentioned earlier and Privoxy which is a web proxy for modifying and filtering webpages. Following these projects ideas, A server which sits in-between the clients and the internet could modify pages achieving a similar solution.

The advantages of this approach would be the privacy of people using the system, as the system is no longer shipping the contents of the user’s webpages to an external server. If the server were to collect logs or was compromised this system could be used as another data point for the tracking and identification of the advertisement servers thus nullifying the purpose of the system this new implantation would mitigate this issue. Another advantage would be the mitigation of browser fingerprinting. This is the process of defining who you are by scraping the setup of your browser like the operating system you are running, the browser and its version, the language and time zone, plugins and many more. This new implementation would effectively make everyone on the network identical.

The biggest disadvantage of this approach is the current resources necessary to run the algorithms. The image classification portion of the system has been well optimised and is effective at classifying images but word2vec relies on training a new dataset on each query this takes considerable resources, making the required hardware expensive. As implementations of this algorithm get more efficient this implementation will become feasible to implement in home environments . Another problem with this approach is by default it won’t work with https traffic; this would have to be solved to make it an effective solution.

#### Removing advertisements

Now that the system can identify images that it thinks are advertisements the next logical step would be to remove them in some fashion. In its current iteration, this can be done with JavaScript by removing the element from the DOM. The implementation proposed above would remove the advertisements before it even gets to the client computer.

#### Word clouds

Word clouds are a simple way of visualising the frequency of a word in text. Early in the development of this project the frequency of the words was disregarded as a method to define what the text is about as in testing the top words tended to not describe the text was about and as such the system uses word2vec which takes every word in the text into account. But when it comes to text like blogs, tutorials, speeches, etc this method of text classification becomes much more viable. The method is also far more lightweight than word2vec meaning it could be implemented in browser either replacing word2vec on sites that contain blog in the URL or bolstering word2vec’s results on other sites.

# Bibliography

1. Manifest V3 [Internet]. Google Docs. [cited 2019 Sep 29]. Available from: https://docs.google.com/document/d/1nPu6Wy4LWR66EFLeYInl3NzzhHzc-qnk4w4PX-0XMw8/edit?usp=embed\_facebook

2. 896897 - Extensions: Implement Manifest V3 - chromium - An open-source project to help move the web forward. - Monorail [Internet]. [cited 2019 Dec 7]. Available from: https://bugs.chromium.org/p/chromium/issues/detail?id=896897&desc=2#c23

3. General Data Protection Regulation (GDPR) – Official Legal Text [Internet]. General Data Protection Regulation (GDPR). [cited 2019 Sep 29]. Available from: https://gdpr-info.eu/

4. Hill R. gorhill/uBlock [Internet]. 2019 [cited 2019 Dec 7]. Available from: https://github.com/gorhill/uBlock

5. AdBlock. AdBlock [Internet]. [cited 2019 Dec 5]. Available from: https://getadblock.com

6. Adblock Plus | The world’s # 1 free ad blocker [Internet]. [cited 2019 Dec 7]. Available from: https://adblockplus.org/en/

7. telekrmor. Home [Internet]. Pi-hole®: A black hole for Internet advertisements. [cited 2019 Dec 7]. Available from: https://pi-hole.net/

8. Google Chrome - The New Chrome & Most Secure Web Browser [Internet]. [cited 2019 Dec 7]. Available from: https://www.google.com/chrome/

9. W3Counter: Global Web Stats - November 2019 [Internet]. [cited 2019 Dec 5]. Available from: https://www.w3counter.com/globalstats.php?year=2019&month=11

10. Firefox Privacy Promise [Internet]. Mozilla. [cited 2019 Dec 7]. Available from: https://www.mozilla.org/en-US/firefox/privacy/

11. Rong X. word2vec Parameter Learning Explained. arXiv:14112738 [cs] [Internet]. 2016 Jun 5 [cited 2019 Dec 5]; Available from: http://arxiv.org/abs/1411.2738

12. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. arXiv:151200567 [cs] [Internet]. 2015 Dec 11 [cited 2019 Dec 5]; Available from: http://arxiv.org/abs/1512.00567

13. WordNet | A Lexical Database for English [Internet]. [cited 2019 Dec 7]. Available from: https://wordnet.princeton.edu/

14. TensorFlow Lite [Internet]. [cited 2019 Dec 7]. Available from: https://www.tensorflow.org/lite

15. Flask-RESTful — Flask-RESTful 0.3.7 documentation [Internet]. [cited 2019 Dec 7]. Available from: https://flask-restful.readthedocs.io/en/latest/

16. EasyList - Overview [Internet]. [cited 2019 Dec 7]. Available from: https://easylist.to/

17. Software Engineering | Classical Waterfall Model [Internet]. GeeksforGeeks. 2018 [cited 2019 Dec 7]. Available from: https://www.geeksforgeeks.org/software-engineering-classical-waterfall-model/

18. Rapid Application Development: Definition, Steps, Advantages and Case Study [Internet]. 2018 [cited 2019 Dec 7]. Available from: https://kissflow.com/rad/rapid-application-development/

19. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed Representations of Words and Phrases and their Compositionality. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, editors. Advances in Neural Information Processing Systems 26 [Internet]. Curran Associates, Inc.; 2013 [cited 2019 Dec 7]. p. 3111–3119. Available from: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf