

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 highlights 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 from ad blocking software, as a result more ads are slipping through the cracks. This system would get around these issues by not having 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*

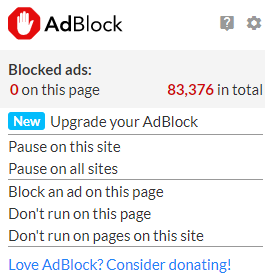
An efficient blocker: easy on memory and CPU footprint, and yet can load and enforce thousands more filters than other popular blockers out there (4). This is a complex interface for an app of this type. This is because the domains of each ad that are blocked are displayed. Control is limited to an on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



*Figure 1- uBlock origin interface*

*AdBlock*

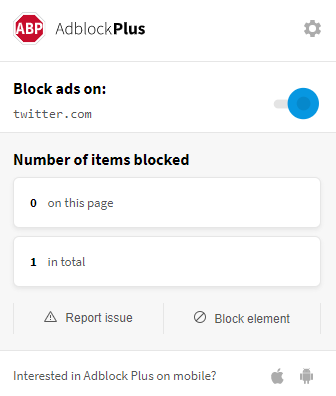
AdBlock (5) does exactly what it says on the tin. It supports the mainstream web browsers as well as android and iOS. The user interface displays the least information but allows the users to pause block and control running in one simple page.



*Figure 2- AdBlock interface*

*AdBlock Plus*

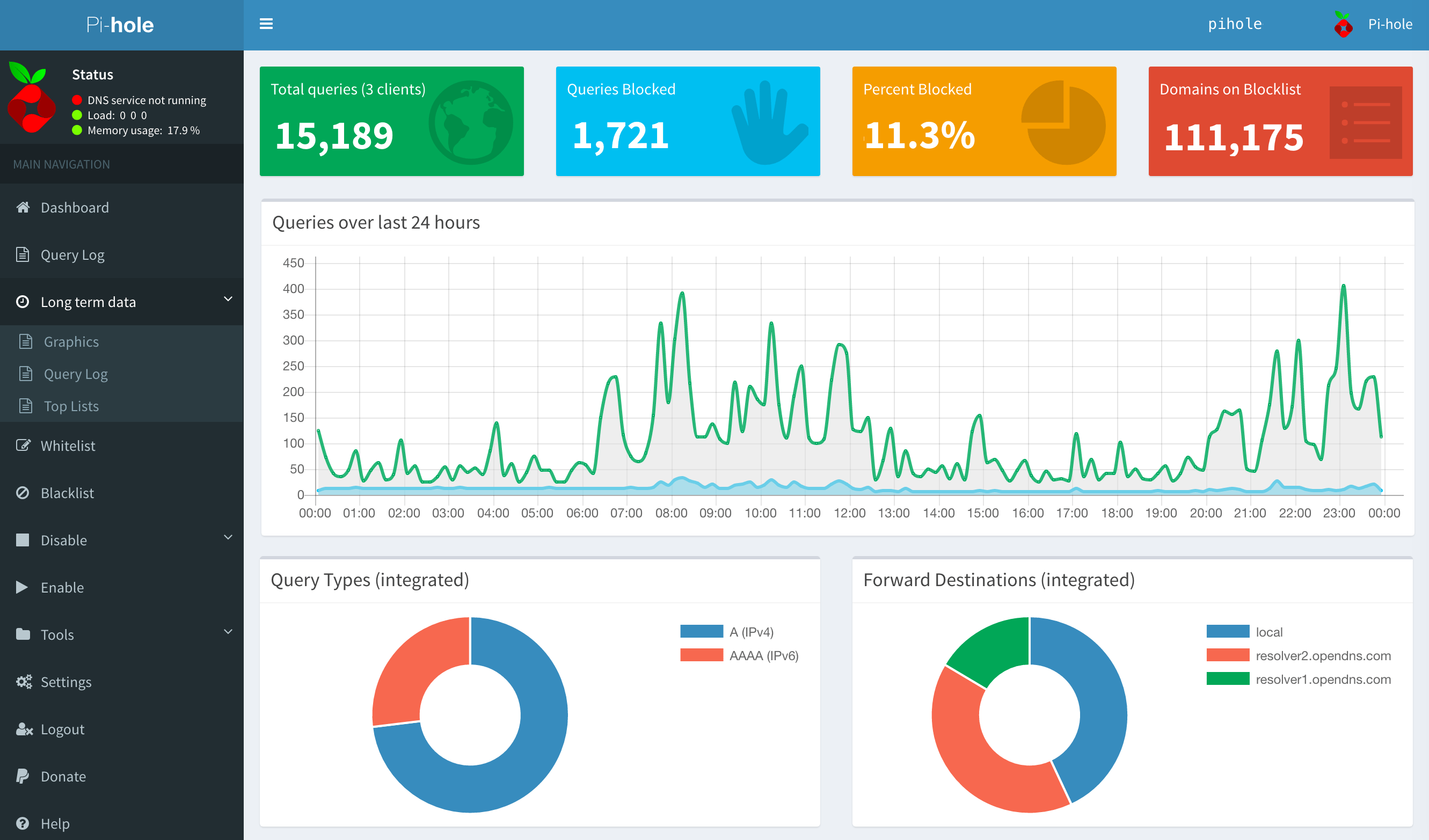
AdBlock plus (6) provides a distraction free browsing experience in the same was as others in this list. AdBlock plus is open-source and gives users access to privacy controls. The interface is very simple, this is a toggle for blocking ads on the current website and a display for advertisements blocked.



*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*

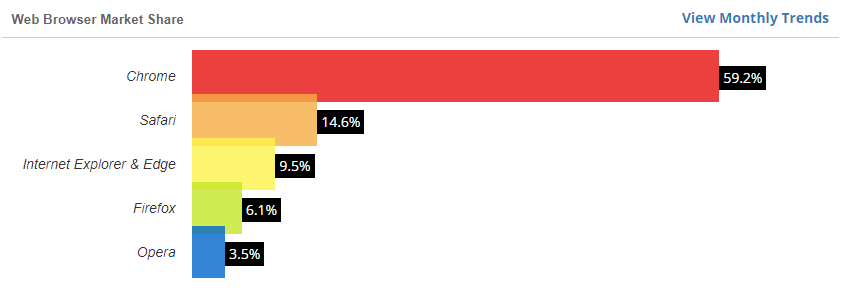
*Conclusion*

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. Theses 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.

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.

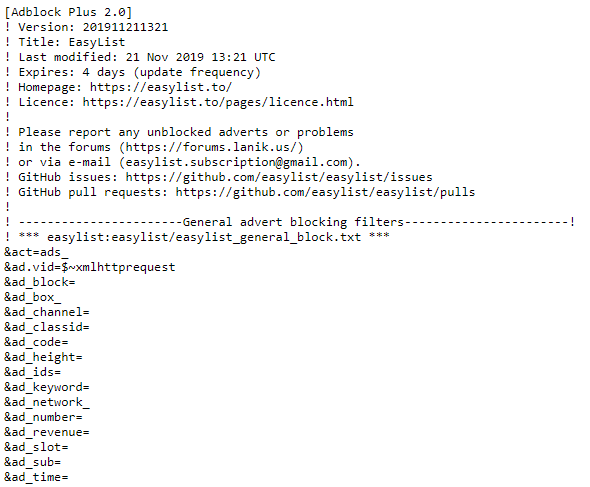
*Conclusion*

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.



## 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.



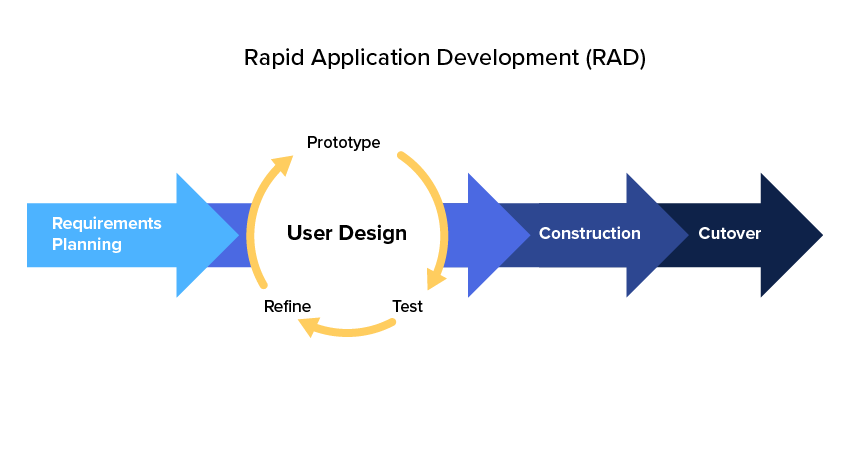
*Figure 7- Waterfall model*

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.

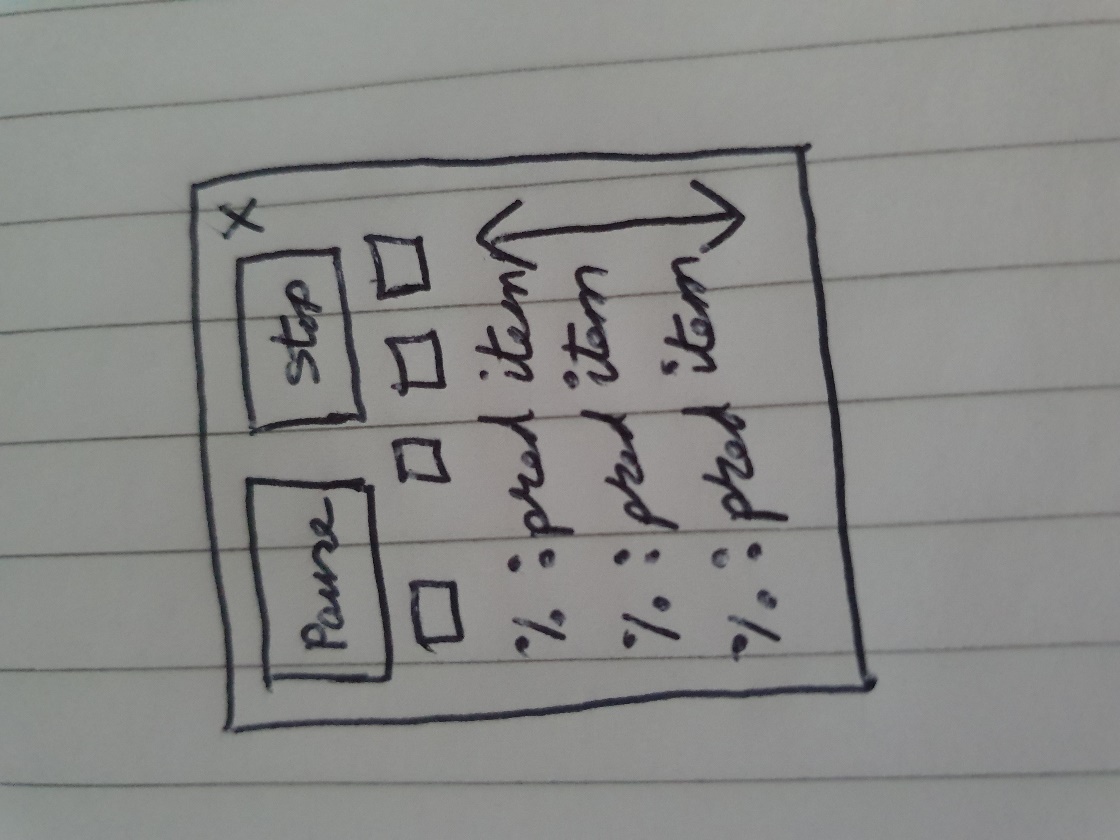
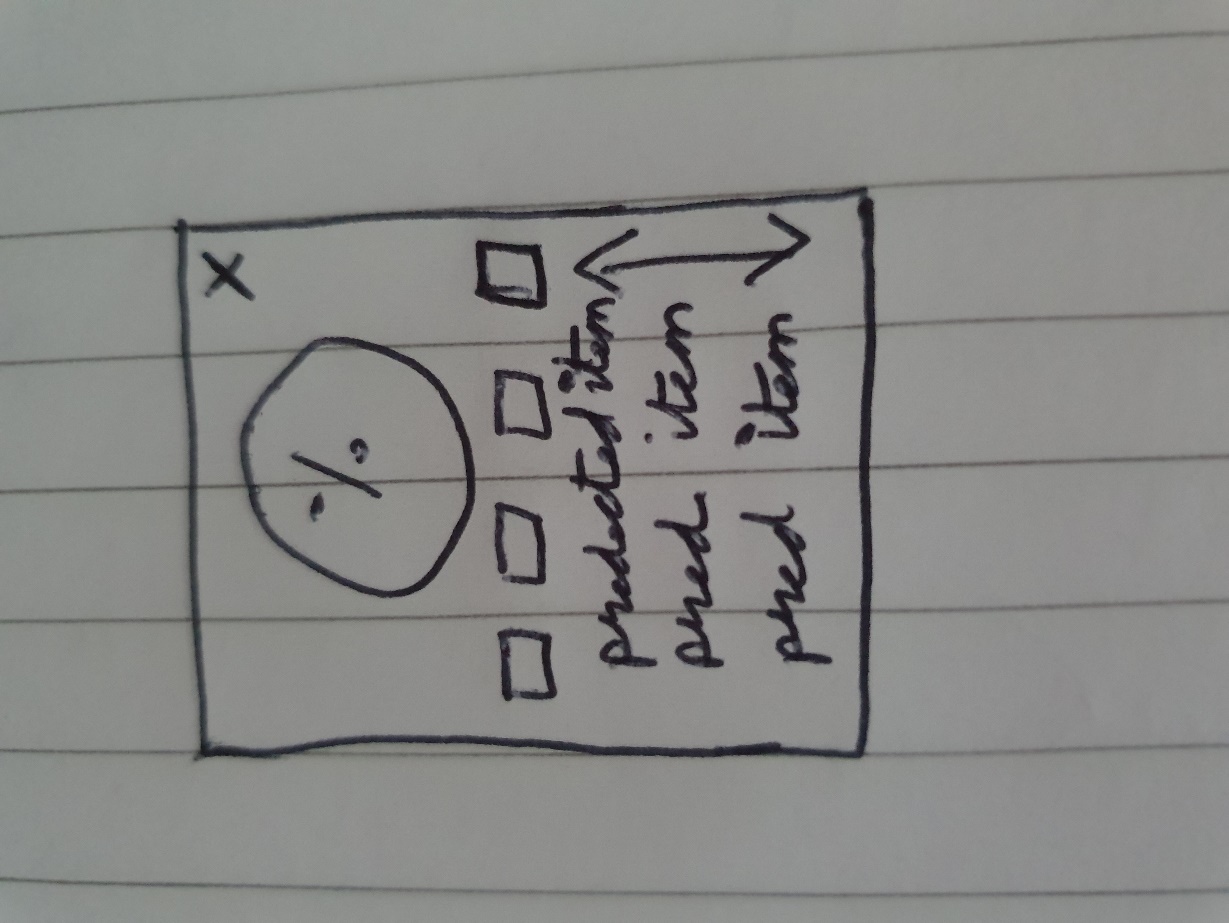
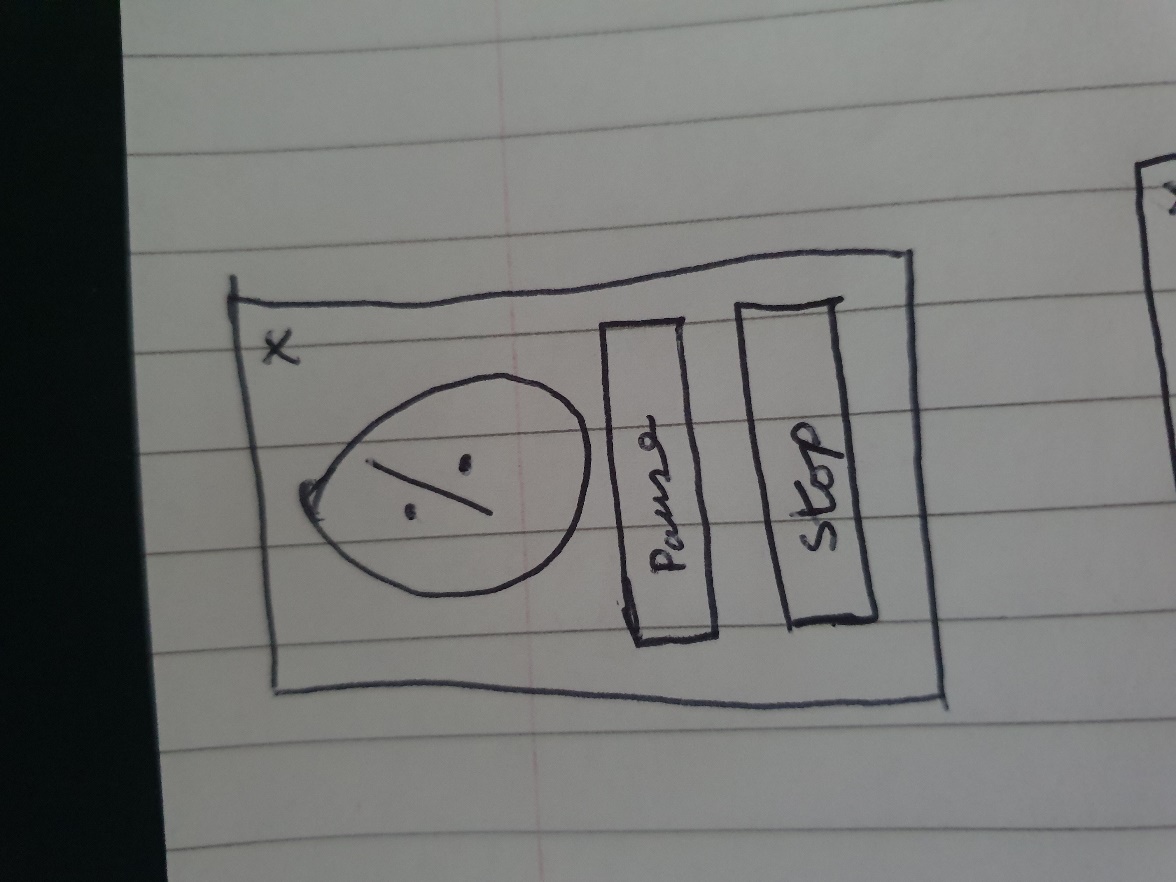
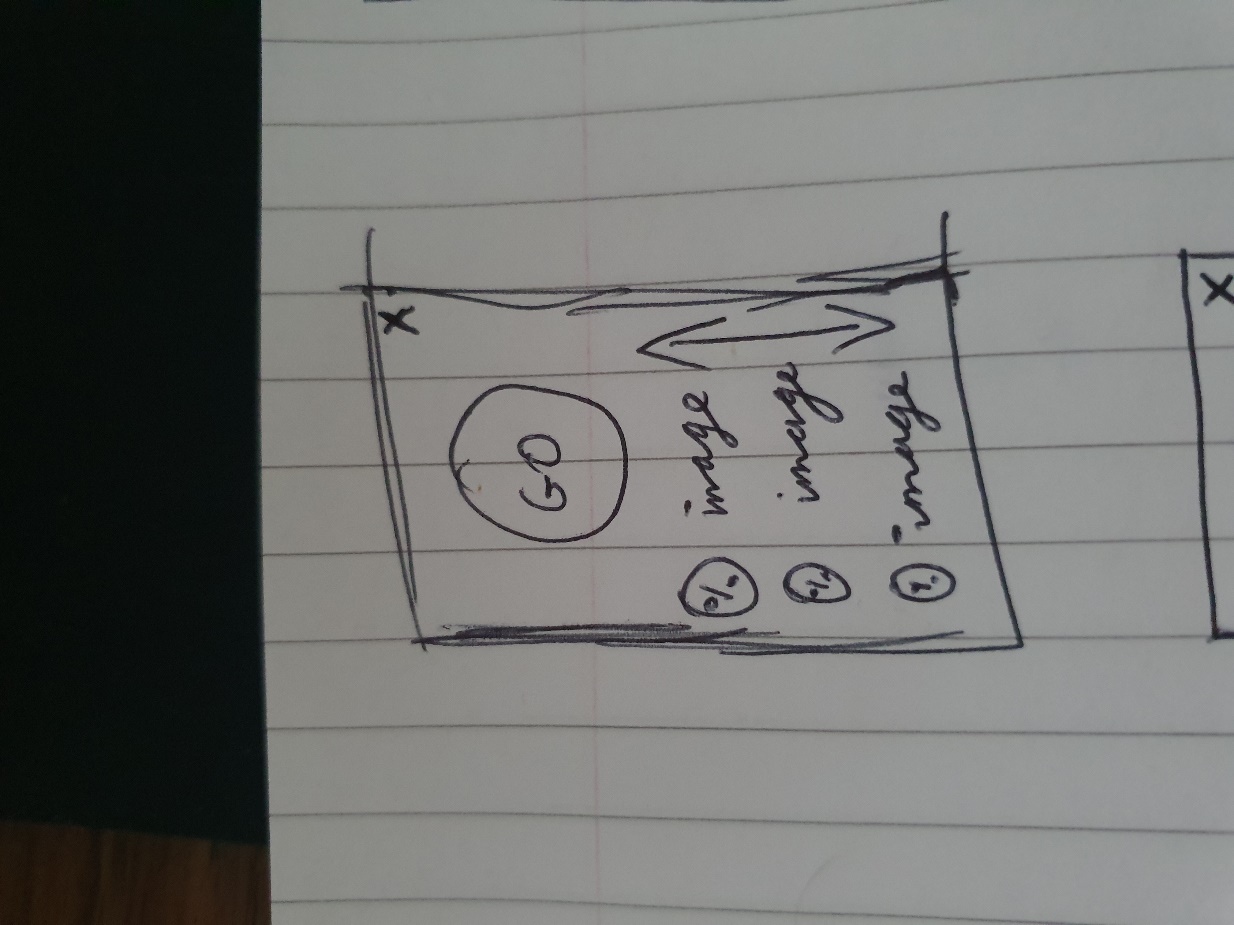
The disadvantages of RAD are mainly based around applying the methodology to larger teams. As this is a solo project this is not a problem. Another disadvantage of this methodology is it is not suitable for larger projects.

## 3.3. Overview of System

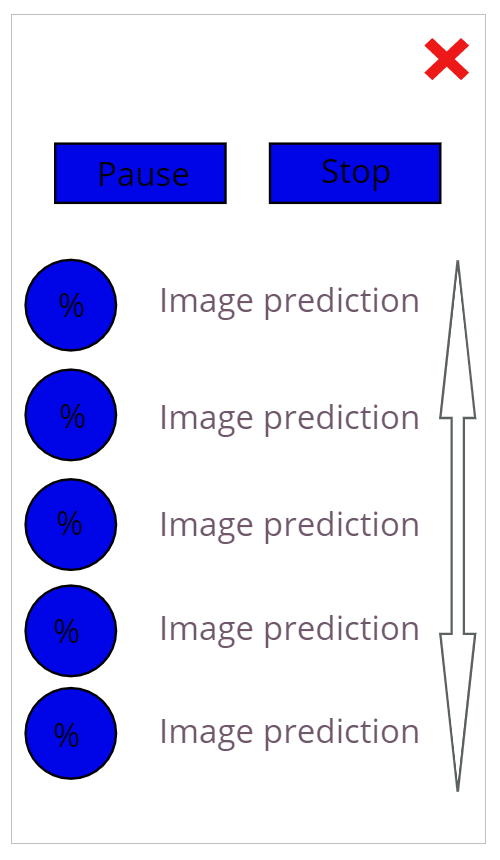
Using the Rapid application development method of software design the first step is define what the system aims to accomplish, this has been completed in part 1.3 of this report. The second stage of RAD is to prototype designs and develop these prototypes. The general approach to the prototype phase 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 has the following ui.

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

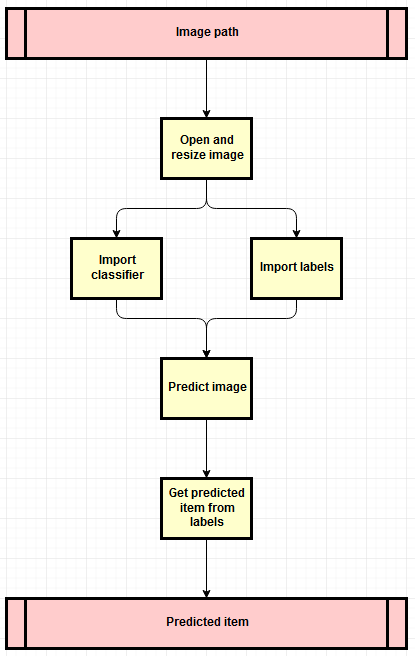
## 3.6. Back-End

The backed will consist of 2 systems, image classification and word embeddings.

*Image classifier*

The image classification system will be based on a retrained inception mode inception. This model is already trained on ImageNet making its predictions already quite accurate, retraining this model for even more common items that we see in advertisements will give an even better result. Training for logos is another option letting us specialize the classifier even more.

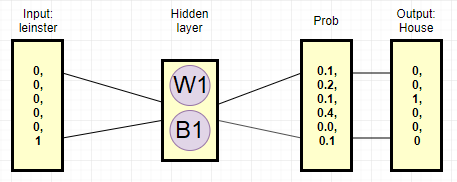
Below is a possible structure for the image prediction. In the final version this will run in browser using TF lite.



*Figure 12 - Image Classification Design*

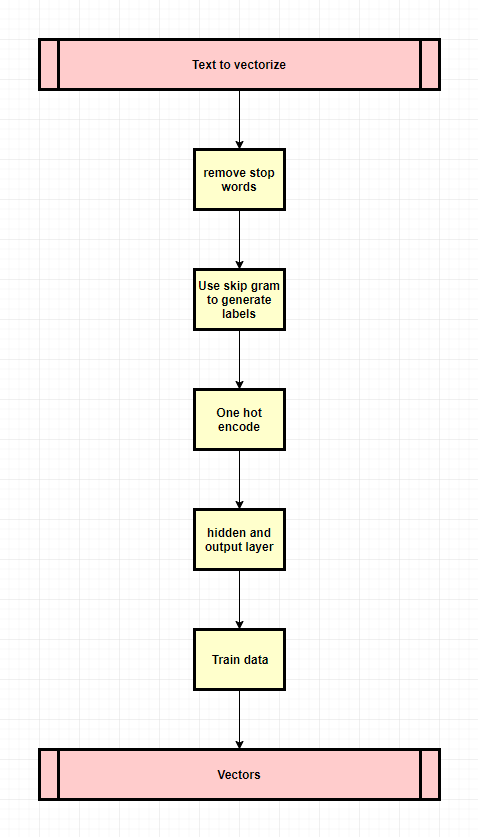
*Word2vec*

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*

## 3.7. Conclusions

# 4. Experiment Development

## 4.1. Introduction

## 4.2. Software Development

## 4.3. Front-End

## 4.4. Middle-Tier

## 4.5. Back-End

## 4.6. Conclusions

# 5. Testing and Evaluation

## 5.1. Introduction

## 5.2. System Testing

## 5.3. System Evaluation

## 5.4. Conclusions

# 6. Conclusions and Future Work

## 6.1. Introduction

## 6.2. Conclusions

## 6.3. Future Work

# Bibliography