David Talan

Fake News Detector

Functional Specification

14387991

30/11/2018

Section 1 – Introduction

1.1 Purpose

This document will provide the functional specification designed to give a synopsis of the 'Fake News Detector'. This will provide a reference for the system design and the intended audience is project coordinator, project supervisor and the CA4000 demonstration examiners.

1.2 Overview

My project idea came from the recent rise in popularity of the term 'fake news', most notably by the U.S. President Donald Trump. Considering this, I decided to create a fake news detector that will be able to analyse a news article and determine if it's of factual nature.

The project will consist of a web application that will take the link of an article and it will analyse it. It will use Natural Language Processing to extract the main contents of the article. I will then use a combination of scrapers and Machine Learning to compare it to existing articles from reliable sources that may be talking about the same topic and see if the initial article coincides with trusted sources.

1.3 Business Context

The web application can be used by anyone or any company who relies on the latest news and current affairs. This would be beneficial to anyone who uses the internet for their source of news and current affairs as it will prevent them from being fed false information and being manipulated.

Section 2 – General Description

2.1 Product/System Functions

The web app will prompt the user to enter a link to the article that they want to be checked. The contents of the article will be obtained by a web scraper and with the use of NLP, the main topics and points of the article will be identified. Using the extracted information, it will then be compared to the other articles from reliable, trusted news sites to test the article's validity. Using Machine Learning, a model will also be trained to analyse the article to determine its validity. The system will then output a score for the article if it's deemed false or factual. If the article is factual, it will return the articles where it found similarities in content that supported its judgement.

2.2 User Characteristics and Objectives

The users of this product will range from young people to the older population. The minimum requirement of these users is to be able to use a computer and navigate their way to the site. Some people might suffer from certain impairments (i.e. visual) but the User Interface will be simple enough and easily understandable for anyone to use. The users are also expected to know how to copy and paste links of their desired article into the text box.

2.3 Operational Scenarios

User pastes in an article link

The user puts in the link to the article and clicks the analyse button to begin the scraping and analysing process. They will be shown a progress bar to give them an idea how long it will take.

User inputs an invalid link

If the user pastes in an invalid link, the system will give them an error message and will prompt them again to input a valid working link.

The system returns the result

The system outputs the score of the article that was inputted, showing the result of the

2.4 Constraints

Time

As the project is ongoing throughout the year, the workload from other modules might take priority in some cases. In semester 1, a lot of research and preparation will be done so the majority of the workload will be ongoing throughout semester 2 up until May. The creation of the Gantt Chart in the final section will definitely help in this case.

Dataset

When it comes to training the Machine Learning model, a lot of data will be needed to train it. Finding sufficient amount of fake news articles data might prove difficult as the topic only became popular recently.

Machine Learning Training

Training the classification model for this project will also prove to have some constraints because of the amount of data that is going to be used. The requirement of a powerful processor might be needed.

Natural Language Processing (NLP)

Since NLP is a topic that haven't been covered in the modules, learning how it functions and applying it the project will add into the time constraint.

Learning new tools for the project

Since this project explores a lot of new areas for me, learning new technologies and suitable tools for the project may prove a challenge.

Section 3 – Functional Requirements

3.1 Retrieving the article

Description

The article's retrieval is the first step in the analysing process. A user inputs the article's link and a scraper will go to the page to extract the content.

Criticality

This process is important as the whole purpose of the fake news detector is to identify said article's credibility as a legitimate news source.

Technical issues

Technical issues may arise if the link the user inputs is invalid. Also if the page they intended to check was deleted, it will return a 404 message, so an exception may need to be in place for cases like that.

Dependencies with other requirements

As the article's retrieval is important to begin the process, a lot of other requirements will be determined by it.

3.2 Extracting the article's content

Description

Using Natural Language Processing, topic segmentation will be applied to the article's text to segment it to certain topics. This will be then used in detecting document similarity in other similar articles online.

Criticality

This is a critical step in the process as it breaks down an articles of varying length down to the most important points that it's trying to convey to the readers. The extraction of the major points will make it easier to compare it to other articles from reliable news sources.

Technical Issues

The NLP step needs to be accurate so that only the relevant points are extracted within the article. If it takes the wrong information from the article, this may lead to a lot of errors and inaccuracies when comparing and applying the classification model on it.

Dependencies with other requirements

The extraction of the content is dependent on 3.1 as it will only work if there is an article to be worked on.

3.3 Comparing to existing articles

Description

With the contents of the target article extracted properly, the content is then compared to existing articles using document similarity.

Criticality

This process plays a major role in identifying fake news. If the extracted contents of the article have been reported by other news sources, it will help solidify the article's and the website's legitimacy as a news source.

Technical issues

The algorithm used for the document similarity needs to work properly to return with a valid, trustworthy result. If the algorithm isn't as accurate as planned, the weights for the combined score with the classification model can be adjusted.

Dependencies with other requirements

The comparison can only happen if the previous requirements are carried out correctly.

3.4 The classification model

Description

The Machine Learning aspect of the project will be the utilisation of a classification model that will analyse the article content. The model will be trained with datasets of fake news articles online. After training, the algorithm will be able to tell if the article is considered fake or real.

Criticality

This is is another essential part of detecting fake news. Similar to the requirement above, it will determine if the article and news site is a legitimate source or not.

Technical issues

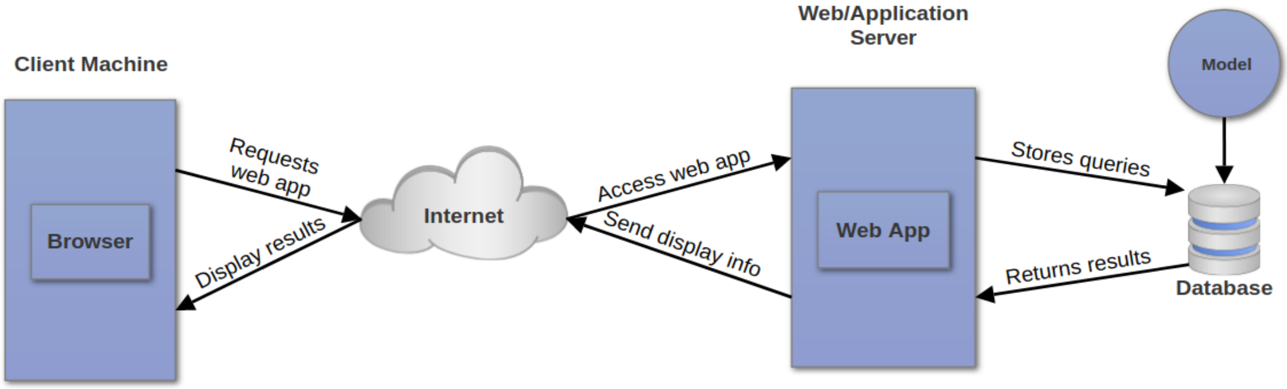
One of the biggest issues will be training the model with datasets. Finding an adequate amount of training data and testing date will be crucial for this part of the project. Otherwise, the classification will be very inaccurate in identifying false information.

Dependencies with other requirements

The model will be dependent on the requirements 3.1 and 3.2 as the algorithm will be applied to extracted article's content. In terms of scoring, the model will be also dependent on how well the document similarity functions. If the document similarity is more accurate than the model, the weight of the score might lean towards that instead of the model, and vice versa.

Section 4 - System Architecture

4.1 Figure A – System Architecture Diagram



Section 5 - High Level Design

Figure 5.1 – Context Diagram

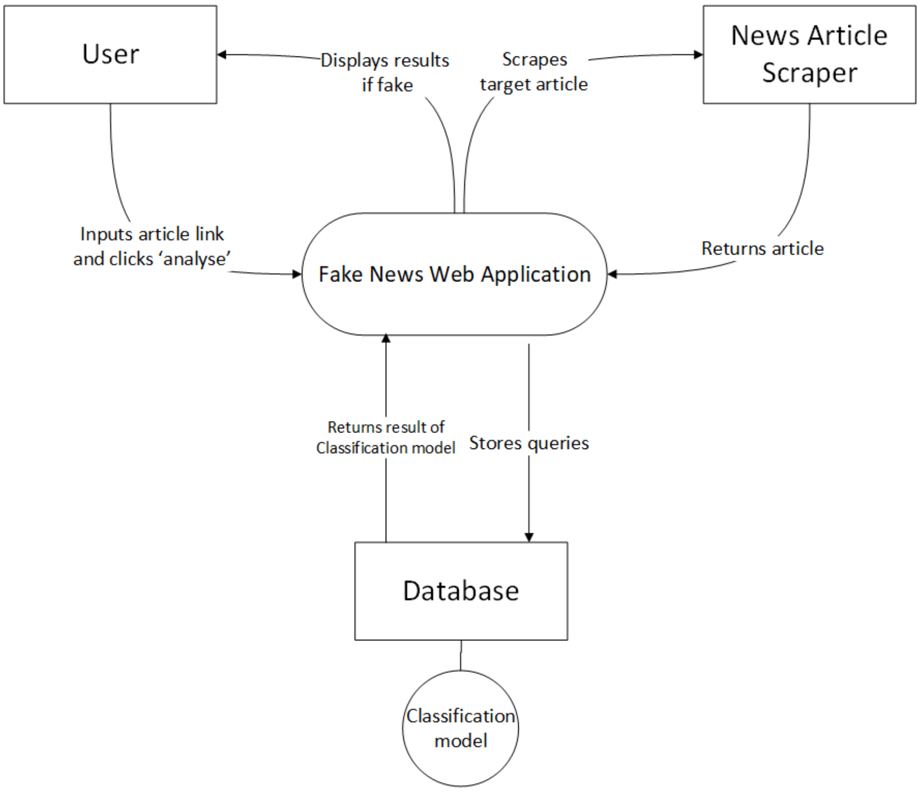
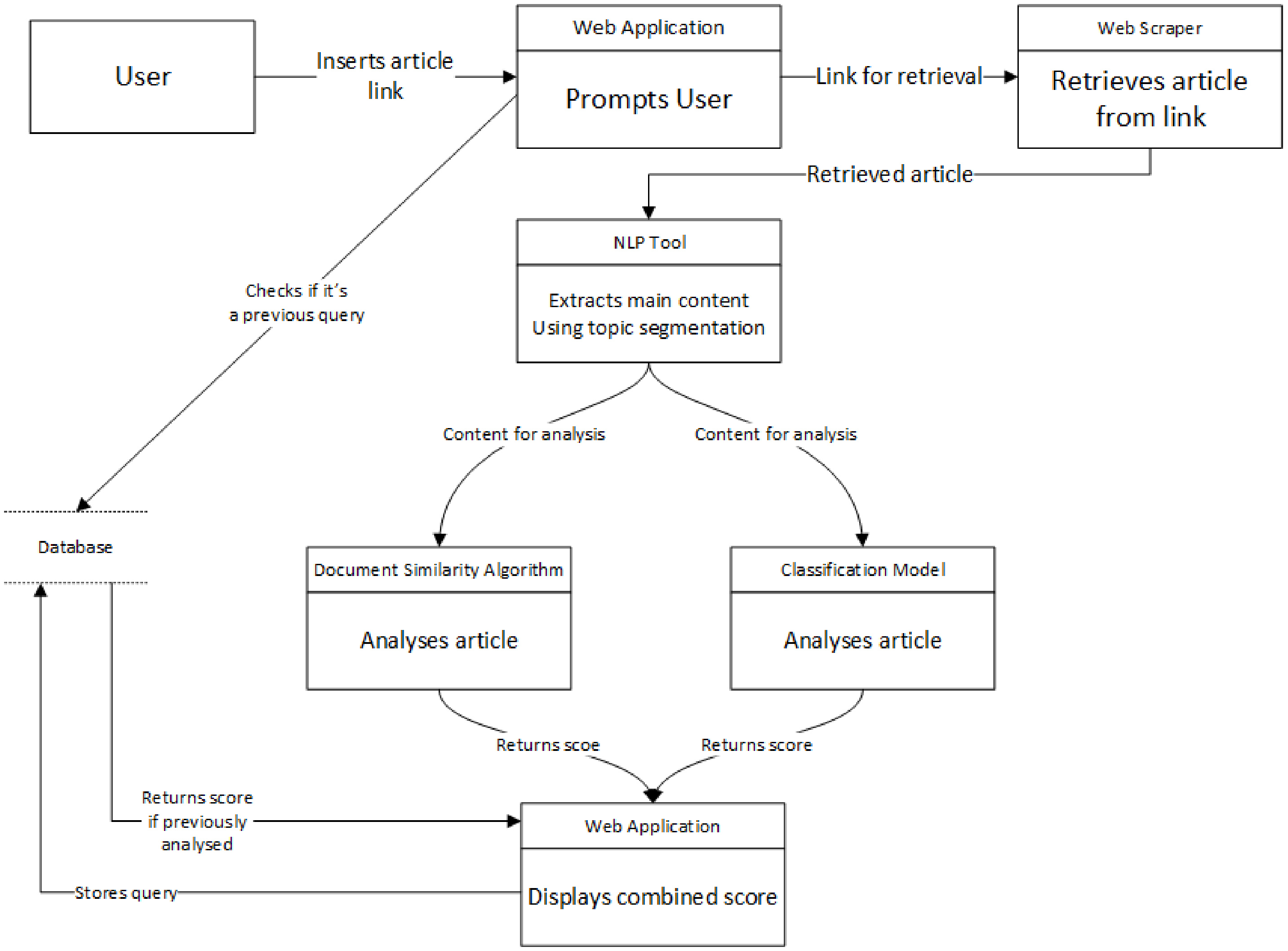
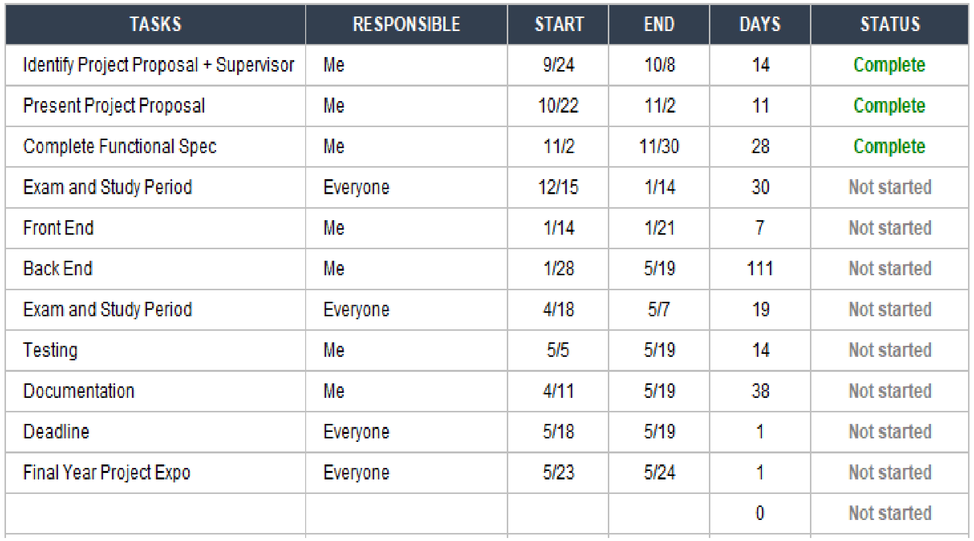


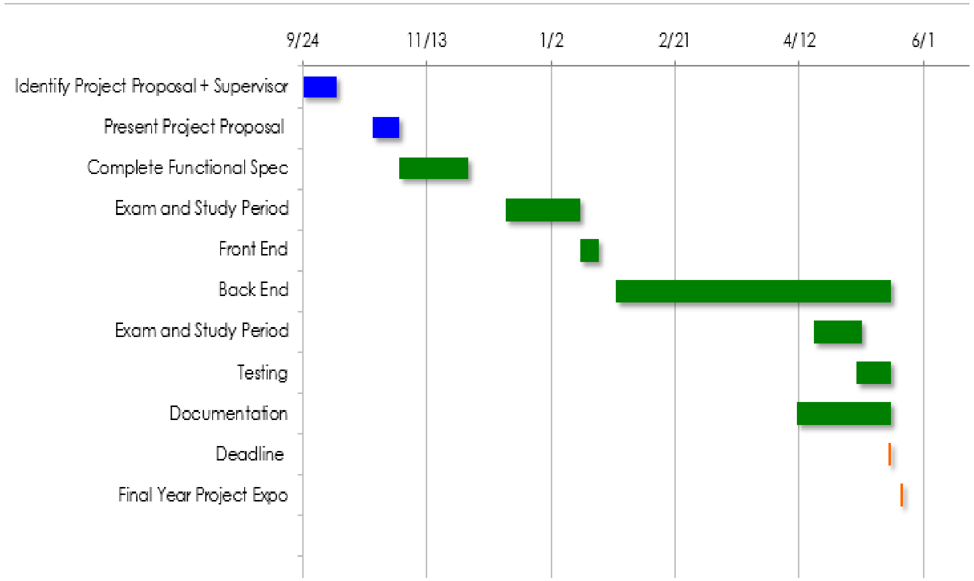
Figure 5.2 – Data Flow Diagram



Section 6 – Preliminary Schedule



TASKS	RESPONSIBLE	START	END	DAYS	STATUS
Identify Project Proposal + Supervisor	Me	9/24	10/8	14	Complete
Present Project Proposal	Me	10/22	11/2	11	Complete
Complete Functional Spec	Me	11/2	11/30	28	Complete
Exam and Study Period	Everyone	12/15	1/14	30	Not started
Front End	Me	1/14	1/21	7	Not started
Back End	Me	1/28	5/19	111	Not started
Exam and Study Period	Everyone	4/18	5/7	19	Not started
Testing	Me	5/5	5/19	14	Not started
Documentation	Me	4/11	5/19	38	Not started
Deadline	Everyone	5/18	5/19	1	Not started
Final Year Project Expo	Everyone	5/23	5/24	1	Not started
				0	Not started



9/24
11/13
1/2
2/21
4/12
6/1
Identify Project Proposal + Supervisor
Present Project Proposal
Complete Functional Spec
Exam and Study Period
Front End
Back End
Exam and Study Period
Testing
Documentation
Deadline
Final Year Project Expo