Sentiment Analysis for Comparing News Articles Across Outlets

Introduction

Machine learning is a branch of artificial intelligence which aims to accurately imitate human learning via statistical methods and large collections of data (Brown, 2021). Within machine learning is a subset called natural language processing which is solely concerned with understanding, extracting and structuring text (Geitgey, 2018). This process involves steps such as segmentation, tokenisation, parts of speech (POS) tagging, lemmatisation and stop-word identification which all work together to break down the text into sentiments, frequencies and word types. This paper makes use of tokenisation and stop-word identification which are both explained in later sections. Sentiment analysis is the application of natural language processing to discover the underlying sentiments in the text (Gupta, 2018). Many analyses are available in this field, but this project focuses on subjectivity and polarity scores which are also explained in later sections. Web scraping involves retrieving and parsing data from online web pages (Amos, 2022). This typically involves using an HTML Parser to split the data into usable components which can be manipulated using programming.

The purpose of this paper is to combine all of these concepts to design a semi-autonomous intelligent agent. An intelligent agent is something that interacts with an environment and makes decisions based on that interaction (*Agents in Artificial Intelligence*, no date). The agent designed in this paper uses an HTML parser to interact with web pages. Based on the data retrieved, it determines how which data to keep and which to discard based on cleaning techniques and the similarity threshold. The resultant action is displaying the retained data in a format which presents information about the articles from each given source, intending to provide the user with a choice between sites which align with their interests differently.

With the end goal of designing a useful application stretching beyond the scope of the project, this will focus on building a web scraper which sifts through news websites for article titles. The program will use text similarity to compare the titles and article contents to determine the likelihood they are written about the same topic. Above a similarity threshold, it will further perform sentiment analysis for each article in each source. Plots will be used to demonstrate the data collected and show which outlets write more positively/negatively and with more or less subjectivity.

Background

This section of the paper aims to provide some background information on how artificial intelligence (AI) and autonomous agents are used in other applications. This should offer some perspective on the prevalence of the two in modern society, but also the wide number of use cases in which they can offer a suitable solution.

Currently, artificial intelligence is divided into many categories and subfields. The forefront of AI research involves expert systems, machine learning, robotics, intelligent decision support systems and pattern recognition (Zhang and Lu, 2021). Within these areas, AI is being applied to a large number of other fields including dentistry (Carrillo-Perez *et al.*, 2022), geotechnical engineering (Baghbani *et al.*, 2022), water treatment (Alam *et al.*, 2022) and metabolic engineering (Jang *et al.*, 2022). All of these publications demonstrate the successes of AI outside of the more obvious and now seemingly generic applications in areas like driverless cars (Manoharan, 2019). Artificial intelligence demonstrably performs well where data is a large focus. Although on a much smaller scale than these papers, this project would not be possible without the collection and preprocessing of data.

Natural language processing stands out on its own with yet another wide array of more specific applications to text and speech processing. It also aids in the medical field (Zhang *et al.*, 2022) and others but perhaps more pertinently in sentiment analysis on website data (Neogi *et al.*, 2021). There is evidence which suggests that web scraping using natural language processing techniques is very popular given the number of research papers and accessible tutorials online (Pascual, 2022). There does however appear to be a gap where scraping and analysing articles between news outlets has not currently been investigated. There are approaches for detecting fake news (Bhutani *et al.*, 2019) and sentiment analyses of articles (Taj, Shaikh and Fatemah Meghji, 2019) but no real comparisons between different sources are present.

Methodology

The following plan was devised to begin moving towards the desired outcomes:

1. Define the environment
    a. Define a list of sources to be used for the model
    b. Investigate which sources are best to be used
2. Define the intelligent agent
    a. Choose a language in which to write the agent
    b. Determine which external modules are necessary for the task
    c. Scrape some articles from each source and store them appropriately
    d. Clean the collected data for analysis
    e. Define a similarity threshold
    f. Define sentiment analysis functions
    g. Compare articles and get polarity and subjectivity scores for each
    h. Plot the results

In the environment in which the agent operates, it can send HTTP requests via the Requests module (*requests*, no date) in Python. This module provides the agent with a response generated from the given web server and allows it to extract the information from a returned object. The agent is not able to make any changes to the environment but simply reacts to the data it is given. This means the agent is of the simple reflex kind but has the potential to upgrade to a model-based reflex or goal-based agent (Choudhuri, 2023). The only interaction to define between the agent and the environment in which sources and the number it accepts. To find the relevant Due to limitations with the project, this is currently fixed but could be made dynamic moving forward, as discussed later in this report.

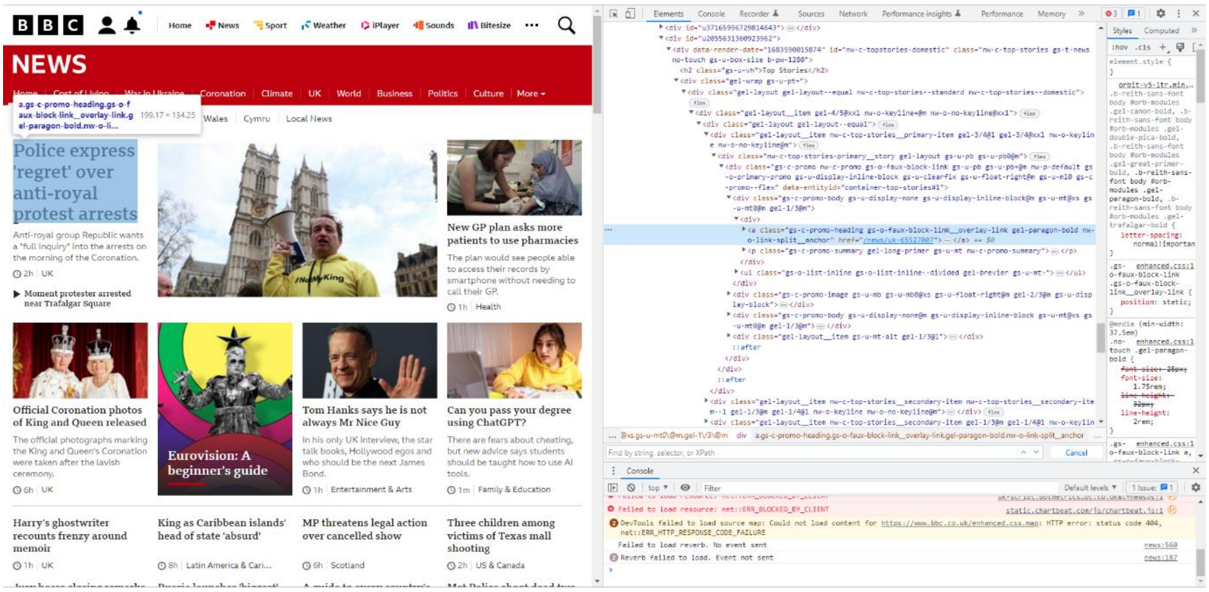
Technically speaking, the agent is a collection of Python files which execute individual tasks semi-automatically. Once the agent (Scraper.py) is run via command prompt, the Cleaner.py, Analyser.py and Plotter.py run subsequently provided no issues in the previous files. If an issue is present, the file will exit and state failure. The agent makes use of the following modules:

- NumPy (*NumPy Reference — NumPy v1.24 Manual*, 2022)
- Pandas (*API reference — pandas 2.0.1 documentation*, no date)

- Requests (*requests*, no date)
- Beautiful Soup (Richardson, no date)
- TextBlob (Loria, no date)
- Natural Language Toolkit (Bird, Klein and Loper, 2009)
- DataFrame_Image (*dataframe_image*, 2023)
- Matplotlib (*matplotlib.pyplot — Matplotlib 3.7.1 documentation*, no date)

Beautiful Soup handles the HTML parsing aspect of the program, meaning it deconstructs the HTML request into elements and allows for storing and manipulation of the data stored within them. This data is then stored in Pandas DataFrames which provide a very simple method for accessing and exporting columns of organised data. The DataFrame_Image module is briefly used to export DataFrames to .png format. The Natural Language Toolkit module is also used for preprocessing the article texts retained for analysis. English stop words, those which do not provide any information on sentiment (Ganesan, no date) are removed from the texts before they are tokenised (Menzli, 2023) and given new values. TextBlob is used to produce subjectivity and polarity scores on the texts. These are a measure of how much opinion and tone play a role in the text. Finally, the Pyplot functions in Matplotlib are used to display and save figures that present data.

Once the above was confirmed, the coding aspect of the project could begin. Working to gather some usable data and clean it for analysis proved tricky and time-consuming initially. This required an understanding of the structure of HTML and which elements provided what. Even then, locating the relevant div class on each page for providing article headings and paragraphs took some time. The final sources used were obtained on the search for the most popular news websites in the UK (Majid, 2023). In this limited scenario, only BBC News and Manchester Evening News were easily usable.



a.gs-c-promo-heading.gs-o-faux-block-link__overlay-link.gel-paragon-bold.nw-o-li... 199.17 × 134.25

Police express 'regret' over anti-royal protest arrests

Anti-royal group Republic wants a "full inquiry" into the arrests on the morning of the Coronation.

2h | UK

Moment protester arrested near Trafalgar Square

New GP plan asks more patients to use pharmacies

The plan would see people able to access their records by smartphone without needing to call their GP.

1h | Health

Official Coronation photos of King and Queen released

The official photographs marking the King and Queen's Coronation were taken after the lavish ceremony.

6h | UK

Eurovision: A beginner's guide

Tom Hanks says he is not always Mr Nice Guy

In his only UK interview, the star talk books, Hollywood egos and who should be the next James Bond.

1h | Entertainment & Arts

Can you pass your degree using ChatGPT?

There are fears about cheating, but new advice says students should be taught how to use AI tools.

1m | Family & Education

Harry's ghostwriter recounts frenzy around memoir

1h | UK

King as Caribbean islands' head of state 'absurd'

8h | Latin America & Cari...

MP threatens legal action over cancelled show

6h | Scotland

Three children among victims of Texas mall shooting

2h | US & Canada

```
▶<div id="u3716599672981464З">…</div>
▼<div id="u2055631360923962">
  ▼<div data-render-date="1683590015874" id="nw-c-topstories-domestic" class="nw-c-top-stories gs-t-news
  no-touch gs-u-box-size b-pw-1280">
     <h2 class="gs-u-vh">Top Stories</h2>
    ▼<div class="gel-wrap gs-u-pt+">
      ▼<div class="gel-layout gel-layout--equal nw-c-top-stories--standard nw-c-top-stories--domestic">
        flex
        ▼<div class="gel-layout__item gel-4/5@xxl nw-o-keyline+@m nw-o-no-keyline@xxl"> flex
          ▼<div class="gel-layout gel-layout--equal"> flex
            ▼<div class="gel-layout__item nw-c-top-stories__primary-item gel-3/4@l gel-3/4@xxl nw-o-keylin
            e nw-o-no-keyline@m"> flex
              ▼<div class="nw-c-top-stories-primary__story gel-layout gs-u-pb gs-u-pb0@m"> flex
                ▼<div class="gs-c-promo nw-c-promo gs-o-faux-block-link gs-u-pb gs-u-pb+@m nw-p-default gs
                -o-primary-promo gs-u-display-inline-block gs-u-clearfix gs-u-float-right@m gs-u-ml0 gs-c
                -promo--flex" data-entityid="container-top-stories#1">
                  ▼<div class="gs-c-promo-body gs-u-display-none gs-u-display-inline-block@m gs-u-mt@xs gs
                  -u-mt0@m gel-1/3@m">
                    ▼<div>
…                     ▶<a class="gs-c-promo-heading gs-o-faux-block-link__overlay-link gel-paragon-bold nw-
                      o-link-split__anchor" href="/news/uk-65527087">…</a> == $0
                      ▶<p class="gs-c-promo-summary gel-long-primer gs-u-mt nw-c-promo-summary">…</p>
                     </div>
                    ▶<ul class="gs-o-list-inline gs-o-list-inline--divided gel-brevier gs-u-mt-">…</ul>
                   </div>
                  ▶<div class="gs-c-promo-image gs-u-mb gs-u-mb0@xs gs-u-float-right@m gel-2/3@m gs-u-disp
                  lay-block">…</div>
                  ▶<div class="gs-c-promo-body gs-u-display-none@m gs-u-display-inline-block gs-u-mt@xs gs
                  -u-mt0@m gel-1/3@m">…</div>
                  ▶<div class="gel-layout__item gs-u-mt-alt gel-1/3@l">…</div>
                   ::after
                 </div>
               </div>
               ::after
             </div>
            ▶<div class="gel-layout__item nw-c-top-stories__secondary-item nw-c-top-stories__secondary-ite
            m--1 gel-1/3@m gel-1/4@l nw-o-keyline nw-o-no-keyline@m">…</div> flex
            ▶<div class="gel-layout__item nw-c-top-stories__secondary-item gel-1/3@m gel-1/4@l nw-o-keylin
… @xs.gs-u-mt0\@m.gel-1\/3\@m  div  a.gs-c-promo-heading.gs-o-faux-block-link__overlay-link.gel-paragon-bold.nw-o-link-split__anchor …
```

```
element.style {
}
.gs-      enhanced.css:1
o-faux-block-link
.gs-o-faux-block-
link__overlay-link {
  position: static;
}
@media (min-width:
37.5em)
.no-      enhanced.css:1
touch .gel-paragon-
bold {
  ~~font-size: 28px;~~
  font-size:
    1.75rem;
  ~~line-height:
    32px;~~
  line-height:
    2rem;
}
```

```
⊗ Failed to load resource: net::ERR_BLOCKED_BY_CLIENT            static.chartbeat.com/js/chartbeat.js:1
2 DevTools failed to load source map: Could not load content for https://www.bbc.co.uk/enhanced.css.map: HTTP error: status code 404,
  net::ERR_HTTP_RESPONSE_CODE_FAILURE
  Failed to load reverb. No event sent                                                                    news:560
2 Reverb failed to load. Event not sent                                                                   news:187
>
```

Figure 1. A demonstration of locating the gs-c-promo-heading div class in Chrome Inspector

Cleaning the data involved removing duplicate article texts along with any issues that arose, such as issues with login blocks and webpage errors. The next step was to define a similarity threshold. This was to be at least 51% similar but would ideally be more in the range of 80-90% similarity to ensure the articles are legitimately on the same topic. In this project, this is calculated using the Jaccard Index (Karabiber, no date). This method of scoring compares two sets of texts, taking the intersection of the two over the union. This method is a rudimentary analysis of similarity since it takes the length of the texts into account rather than the sentiment but for this project is appropriate in the interest of building foundations.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Figure 2. Calculating the Jaccard Index (Facer, 2018)

Finally, before plotting the results, the functions for defining subjectivity and polarity scores are created. Fortunately, this is very easily accessible in a TextBlob object which takes the text as an argument. From here these scores are accessible as variables in the sentiment function. Subjectivity is set as a decimal in the range of 0-1 where zero denotes no subjectivity and 1 denotes complete subjectivity. Similarly, polarity is within (–1)-1 where -1 denotes complete negativity and 1 denotes complete positivity (Loria, no date).

Questions, Experiments and General Testing

The experiments set out in this project are inherently tied to the more prominent questions that demonstrate whether the coded solution is both practical and successful. Measuring the repeatability of the experiments from a statistical standpoint was somewhat difficult due to the nature of the project. There were not many opportunities to obtain repeatable statistics due to articles changing daily on news websites but the routines in the code have been run multiple times to prove the solution achieves the desired outcome without producing errors or throwing exceptions. Rather than exclusively measuring statistics, more focus has been put on dividing the functionalities and observing the result at each stage. The questions set out for this project are:

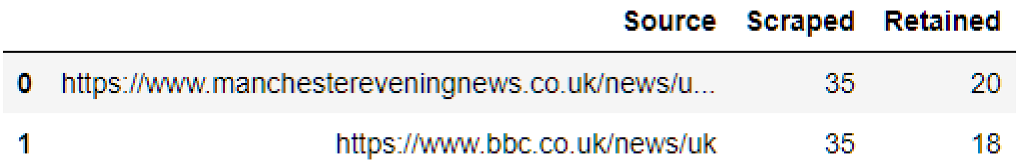
- Does the program successfully scrape news outlet websites for information?
- Does the program successfully rearrange and clean the obtained data?
- Does the program successfully make use of natural language processing?
- Does the program successfully provide polarity and subjectivity scores for articles over a similarity threshold?
- Are the final results expected or indicative of room for improvement?

When running experiments to answer these questions, the evidence of success is found in multiple steps along the way. First of all, the tasks associated with the project have been divided into different Python files, although all run from the original file, emulating the idea of an autonomous agent. Each file, Scraper.py, Cleaner.py, Analyser.py and Plotter.py make use of try, except blocks (*Try and Except in Python*, no date) and print statements to determine at the highest level whether the code runs without any issues. If the file runs without any errors, Python will print output a statement to the console stating its success. If the file fails to run for whatever reason, Python will do the same stating failure. This is a very rudimentary use of this kind of error handling but enough to prove the bot follows through from start to finish.

Further to this, at any stage in which the data is changed or handled in any way, the results are output to either a .csv or .png file. This allows for very easy comparisons between steps to measure whether the correct changes have been applied. In the results section, these outputs demonstrate the last run of the code in order of steps with a discussion at each step.

Final Results

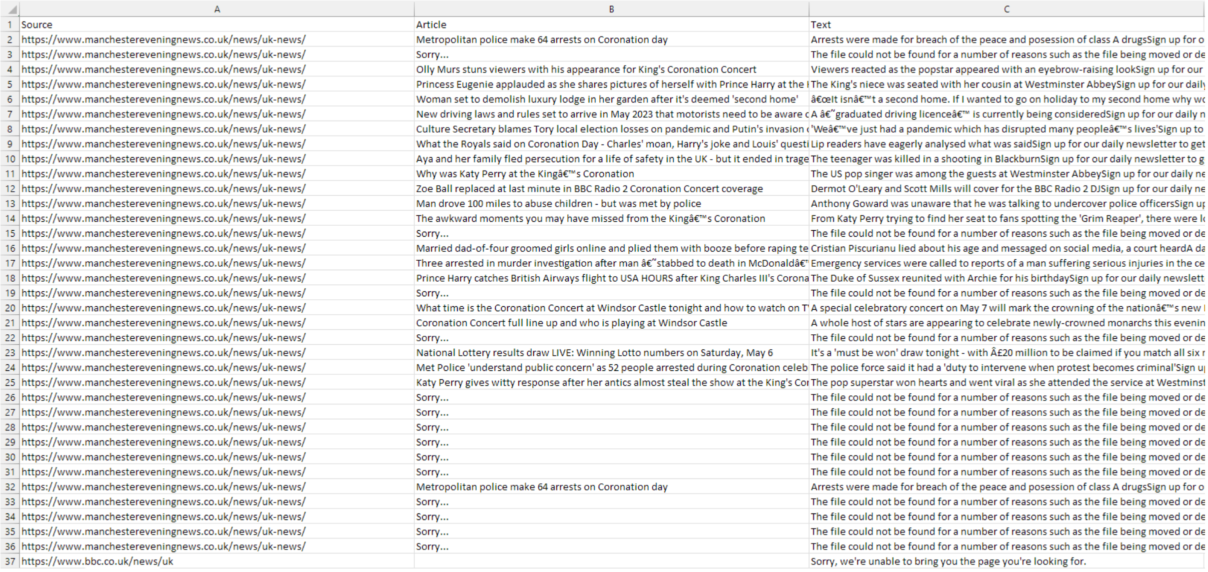
The first step in the testing process is to determine whether the web scraper portion of the program successfully retrieves and cleans the data obtained. This is measured by outputting a Pandas DataFrame to a .png and observing the source, the number of scraped articles and the number retained after removing any duplicate articles as well as those with blank values or any issues with login information or authorisations. Figure 3 demonstrates that the program has successfully scraped 35 articles but only 20 and 18 of them were retained from Manchester Evening News and BBC News respectively.



	Source	Scraped	Retained
0	https://www.manchestereveningnews.co.uk/news/u...	35	20
1	https://www.bbc.co.uk/news/uk	35	18

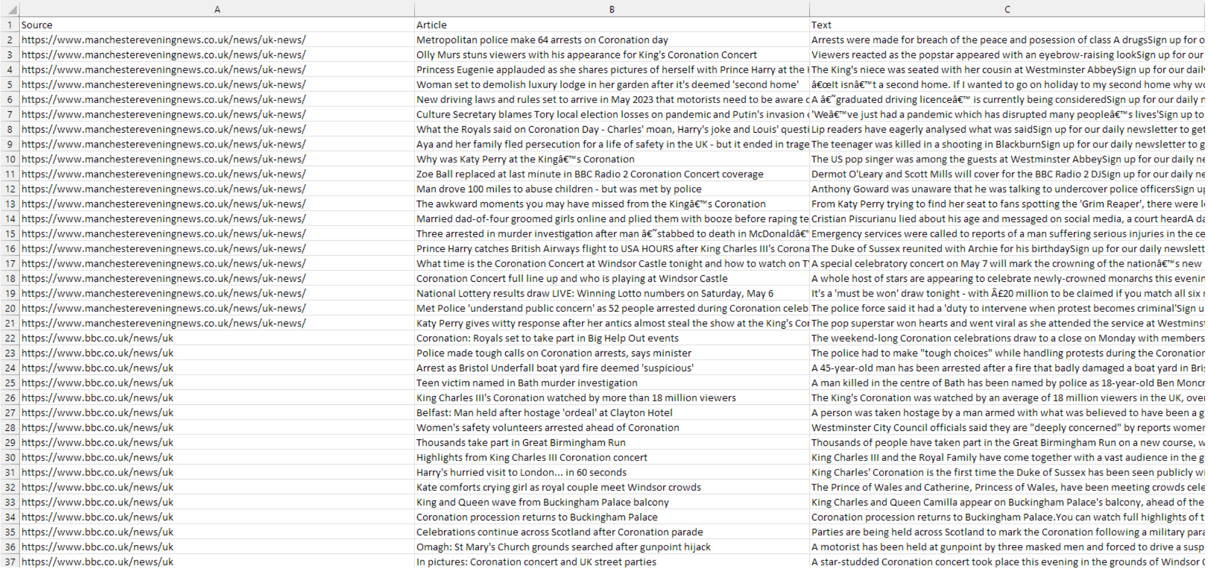
Figure 3. The retained number of articles from each source after scraping and cleaning

After noting the number of articles that are retained for analysis, the retrieved data can be observed as it was when pulled from the web server and after the program has removed the data that is deemed unusable for the project. Figure 4 shows that within the number of total articles scraped, many of them have issues related to finding the associated files.



	A	B	C
1	Source	Article	Text
2	https://www.manchestereveningnews.co.uk/news/uk-news/	Metropolitan police make 64 arrests on Coronation day	Arrests were made for breach of the peace and posession of class A drugsSign up for o
3	https://www.manchestereveningnews.co.uk/news/uk-news/	Sorry...	The file could not be found for a number of reasons such as the file being moved or de
4	https://www.manchestereveningnews.co.uk/news/uk-news/	Olly Murs stuns viewers with his appearance for King's Coronation Concert	Viewers reacted as the popstar appeared with an eyebrow-raising lookSign up for our
5	https://www.manchestereveningnews.co.uk/news/uk-news/	Princess Eugenie applauded as she shares pictures of herself with Prince Harry at the	The King's niece was seated with her cousin at Westminster AbbeySign up for our daily
6	https://www.manchestereveningnews.co.uk/news/uk-news/	Woman set to demolish luxury lodge in her garden after it's deemed 'second home'	â€œIt isnâ€™t a second home. If I wanted to go on holiday to my second home why wo
7	https://www.manchestereveningnews.co.uk/news/uk-news/	New driving laws and rules set to arrive in May 2023 that motorists need to be aware c	A â€˜graduated driving licenceâ€™ is currently being consideredSign up for our daily n
8	https://www.manchestereveningnews.co.uk/news/uk-news/	Culture Secretary blames Tory local election losses on pandemic and Putin's invasion (	'Weâ€™ve just had a pandemic which has disrupted many peopleâ€™s lives'Sign up to
9	https://www.manchestereveningnews.co.uk/news/uk-news/	What the Royals said on Coronation Day - Charles' moan, Harry's joke and Louis' questi	Lip readers have eagerly analysed what was saidSign up for our daily newsletter to get
10	https://www.manchestereveningnews.co.uk/news/uk-news/	Aya and her family fled persecution for a life of safety in the UK - but it ended in trage	The teenager was killed in a shooting in BlackburnSign up for our daily newsletter to g
11	https://www.manchestereveningnews.co.uk/news/uk-news/	Why was Katy Perry at the Kingâ€™s Coronation	The US pop singer was among the guests at Westminster AbbeySign up for our daily ne
12	https://www.manchestereveningnews.co.uk/news/uk-news/	Zoe Ball replaced at last minute in BBC Radio 2 Coronation Concert coverage	Dermot O'Leary and Scott Mills will cover for the BBC Radio 2 DJSign up for our daily ne
13	https://www.manchestereveningnews.co.uk/news/uk-news/	Man drove 100 miles to abuse children - but was met by police	Anthony Goward was unaware that he was talking to undercover police officersSign up
14	https://www.manchestereveningnews.co.uk/news/uk-news/	The awkward moments you may have missed from the Kingâ€™s Coronation	From Katy Perry trying to find her seat to fans spotting the 'Grim Reaper', there were lo
15	https://www.manchestereveningnews.co.uk/news/uk-news/	Sorry...	The file could not be found for a number of reasons such as the file being moved or de
16	https://www.manchestereveningnews.co.uk/news/uk-news/	Married dad-of-four groomed girls online and plied them with booze before raping te	Cristian Piscurianu lied about his age and messaged on social media, a court heardA da
17	https://www.manchestereveningnews.co.uk/news/uk-news/	Three arrested in murder investigation after man â€˜stabbed to death in McDonaldâ€™	Emergency services were called to reports of a man suffering serious injuries in the ce
18	https://www.manchestereveningnews.co.uk/news/uk-news/	Prince Harry catches British Airways flight to USA HOURS after King Charles III's Corona	The Duke of Sussex reunited with Archie for his birthdaySign up for our daily newslett
19	https://www.manchestereveningnews.co.uk/news/uk-news/	Sorry...	The file could not be found for a number of reasons such as the file being moved or de
20	https://www.manchestereveningnews.co.uk/news/uk-news/	What time is the Coronation Concert at Windsor Castle tonight and how to watch on T	A special celebratory concert on May 7 will mark the crowning of the nationâ€™s new
21	https://www.manchestereveningnews.co.uk/news/uk-news/	Coronation Concert full line up and who is playing at Windsor Castle	A whole host of stars are appearing to celebrate newly-crowned monarchs this evenin
22	https://www.manchestereveningnews.co.uk/news/uk-news/	Sorry...	The file could not be found for a number of reasons such as the file being moved or de
23	https://www.manchestereveningnews.co.uk/news/uk-news/	National Lottery results draw LIVE: Winning Lotto numbers on Saturday, May 6	It's a 'must be won' draw tonight - with Â£20 million to be claimed if you match all six r
24	https://www.manchestereveningnews.co.uk/news/uk-news/	Met Police 'understand public concern' as 52 people arrested during Coronation celeb	The police force said it had a 'duty to intervene when protest becomes criminal'Sign up
25	https://www.manchestereveningnews.co.uk/news/uk-news/	Katy Perry gives witty response after her antics almost steal the show at the King's Cor	The pop superstar won hearts and went viral as she attended the service at Westminst
26	https://www.manchestereveningnews.co.uk/news/uk-news/	Sorry...	The file could not be found for a number of reasons such as the file being moved or de
27	https://www.manchestereveningnews.co.uk/news/uk-news/	Sorry...	The file could not be found for a number of reasons such as the file being moved or de
28	https://www.manchestereveningnews.co.uk/news/uk-news/	Sorry...	The file could not be found for a number of reasons such as the file being moved or de
29	https://www.manchestereveningnews.co.uk/news/uk-news/	Sorry...	The file could not be found for a number of reasons such as the file being moved or de
30	https://www.manchestereveningnews.co.uk/news/uk-news/	Sorry...	The file could not be found for a number of reasons such as the file being moved or de
31	https://www.manchestereveningnews.co.uk/news/uk-news/	Sorry...	The file could not be found for a number of reasons such as the file being moved or de
32	https://www.manchestereveningnews.co.uk/news/uk-news/	Metropolitan police make 64 arrests on Coronation day	Arrests were made for breach of the peace and posession of class A drugsSign up for o
33	https://www.manchestereveningnews.co.uk/news/uk-news/	Sorry...	The file could not be found for a number of reasons such as the file being moved or de
34	https://www.manchestereveningnews.co.uk/news/uk-news/	Sorry...	The file could not be found for a number of reasons such as the file being moved or de
35	https://www.manchestereveningnews.co.uk/news/uk-news/	Sorry...	The file could not be found for a number of reasons such as the file being moved or de
36	https://www.manchestereveningnews.co.uk/news/uk-news/	Sorry...	The file could not be found for a number of reasons such as the file being moved or de
37	https://www.bbc.co.uk/news/uk		Sorry, we're unable to bring you the page you're looking for.

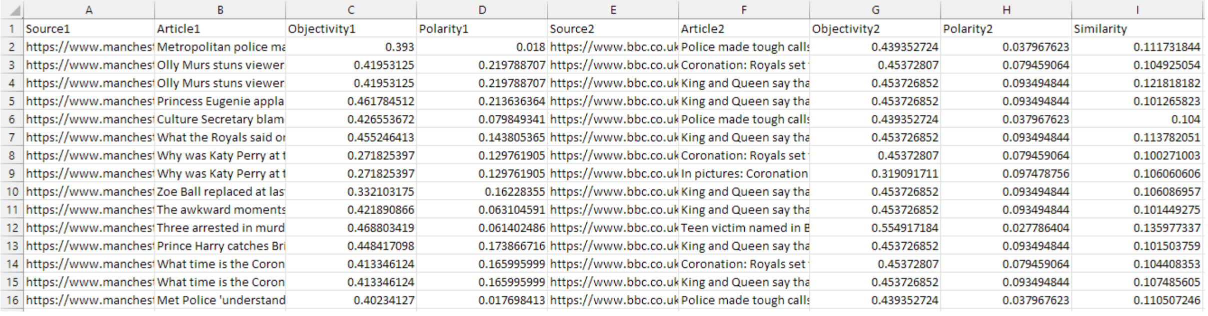
Figure 4. A representation of the ‘dirty’ data .csv file after scraping is complete



Source	Article	Text
https://www.manchestereveningnews.co.uk/news/uk-news/	Metropolitan police make 64 arrests on Coronation day	Arrests were made for breach of the peace and posession of class A drugsSign up for o
https://www.manchestereveningnews.co.uk/news/uk-news/	Olly Murs stuns viewers with his appearance for King's Coronation Concert	Viewers reacted as the popstar appeared with an eyebrow-raising lookSign up for our
https://www.manchestereveningnews.co.uk/news/uk-news/	Princess Eugenie applauded as she shares pictures of herself with Prince Harry at the	The King's niece was seated with her cousin at Westminster AbbeySign up for our daily
https://www.manchestereveningnews.co.uk/news/uk-news/	Woman set to demolish luxury lodge in her garden after it's deemed 'second home'	â€œIt isnâ€™t a second home. If I wanted to go on holiday to my second home why wo
https://www.manchestereveningnews.co.uk/news/uk-news/	New driving laws and rules set to arrive in May 2023 that motorists need to be aware c	A â€˜graduated driving licenceâ€™ is currently being consideredSign up for our daily n
https://www.manchestereveningnews.co.uk/news/uk-news/	Culture Secretary blames Tory local election losses on pandemic and Putin's invasion c	'Weâ€™ve just had a pandemic which has disrupted many peopleâ€™s lives'Sign up to
https://www.manchestereveningnews.co.uk/news/uk-news/	What the Royals said on Coronation Day - Charles' moan, Harry's joke and Louis' questi	Lip readers have eagerly analysed what was saidSign up for our daily newsletter to get
https://www.manchestereveningnews.co.uk/news/uk-news/	Aya and her family fled persecution for a life of safety in the UK - but it ended in trage	The teenager was killed in a shooting in BlackburnSign up for our daily newsletter to g
https://www.manchestereveningnews.co.uk/news/uk-news/	Why was Katy Perry at the Kingâ€™s Coronation	The US pop singer was among the guests at Westminster AbbeySign up for our daily ne
https://www.manchestereveningnews.co.uk/news/uk-news/	Zoe Ball replaced at last minute in BBC Radio 2 Coronation Concert coverage	Dermot O'Leary and Scott Mills will cover for the BBC Radio 2 DJSign up for our daily ne
https://www.manchestereveningnews.co.uk/news/uk-news/	Man drove 100 miles to abuse children - but was met by police	Anthony Goward was unaware that he was talking to undercover police officersSign up
https://www.manchestereveningnews.co.uk/news/uk-news/	The awkward moments you may have missed from the Kingâ€™s Coronation	From Katy Perry trying to find her seat to fans spotting the 'Grim Reaper', there were l
https://www.manchestereveningnews.co.uk/news/uk-news/	Married dad-of-four groomed girls online and plied them with booze before raping te	Cristian Piscurianu lied about his age and messaged on social media, a court heardA da
https://www.manchestereveningnews.co.uk/news/uk-news/	Three arrested in murder investigation after man â€˜stabbed to death in McDonaldâ€™	Emergency services were called to reports of a man suffering serious injuries in the ce
https://www.manchestereveningnews.co.uk/news/uk-news/	Prince Harry catches British Airways flight to USA HOURS after King Charles III's Corona	The Duke of Sussex reunited with Archie for his birthdaySign up for our daily newslett
https://www.manchestereveningnews.co.uk/news/uk-news/	What time is the Coronation Concert at Windsor Castle tonight and how to watch on T	A special celebratory concert on May 7 will mark the crowning of the nationâ€™s new
https://www.manchestereveningnews.co.uk/news/uk-news/	Coronation Concert full line up and who is playing at Windsor Castle	A whole host of stars are appearing to celebrate newly-crowned monarchs this evenin
https://www.manchestereveningnews.co.uk/news/uk-news/	National Lottery results draw LIVE: Winning Lotto numbers on Saturday, May 6	It's a 'must be won' draw tonight - with Â£20 million to be claimed if you match all six
https://www.manchestereveningnews.co.uk/news/uk-news/	Met Police 'understand public concern' as 52 people arrested during Coronation celeb	The police force said it had a 'duty to intervene when protest becomes criminal'Sign u
https://www.manchestereveningnews.co.uk/news/uk-news/	Katy Perry gives witty response after her antics almost steal the show at the King's Cor	The pop superstar won hearts and went viral as she attended the service at Westminst
https://www.bbc.co.uk/news/uk	Coronation: Royals set to take part in Big Help Out events	The weekend-long Coronation celebrations draw to a close on Monday with members
https://www.bbc.co.uk/news/uk	Police made tough calls on Coronation arrests, says minister	The police had to make "tough choices" while handling protests during the Coronatior
https://www.bbc.co.uk/news/uk	Arrest as Bristol Underfall boat yard fire deemed 'suspicious'	A 45-year-old man has been arrested after a fire that badly damaged a boat yard in Bri
https://www.bbc.co.uk/news/uk	Teen victim named in Bath murder investigation	A man killed in the centre of Bath has been named by police as 18-year-old Ben Moncr
https://www.bbc.co.uk/news/uk	King Charles III's Coronation watched by more than 18 million viewers	The King's Coronation was watched by an average of 18 million viewers in the UK, over
https://www.bbc.co.uk/news/uk	Belfast: Man held after hostage 'ordeal' at Clayton Hotel	A person was taken hostage by a man armed with what was believed to have been a g
https://www.bbc.co.uk/news/uk	Women's safety volunteers arrested ahead of Coronation	Westminster City Council officials said they are "deeply concerned" by reports womer
https://www.bbc.co.uk/news/uk	Thousands take part in Great Birmingham Run	Thousands of people have taken part in the Great Birmingham Run on a new course, w
https://www.bbc.co.uk/news/uk	Highlights from King Charles III Coronation concert	King Charles III and the Royal Family have come together with a vast audience in the g
https://www.bbc.co.uk/news/uk	Harry's hurried visit to London... in 60 seconds	King Charles' Coronation is the first time the Duke of Sussex has been seen publicly wi
https://www.bbc.co.uk/news/uk	Kate comforts crying girl as royal couple meet Windsor crowds	The Prince of Wales and Catherine, Princess of Wales, have been meeting crowds cele
https://www.bbc.co.uk/news/uk	King and Queen wave from Buckingham Palace balcony	King Charles and Queen Camilla appear on Buckingham Palace's balcony, ahead of the
https://www.bbc.co.uk/news/uk	Coronation procession returns to Buckingham Palace	Coronation procession returns to Buckingham Palace.You can watch full highlights of t
https://www.bbc.co.uk/news/uk	Celebrations continue across Scotland after Coronation parade	Parties are being held across Scotland to mark the Coronation following a military para
https://www.bbc.co.uk/news/uk	Omagh: St Mary's Church grounds searched after gunpoint hijack	A motorist has been held at gunpoint by three masked men and forced to drive a susp
https://www.bbc.co.uk/news/uk	In pictures: Coronation concert and UK street parties	A star-studded Coronation concert took place this evening in the grounds of Windsor (

Figure 5. A representation of the 'clean' data after removing unusable rows

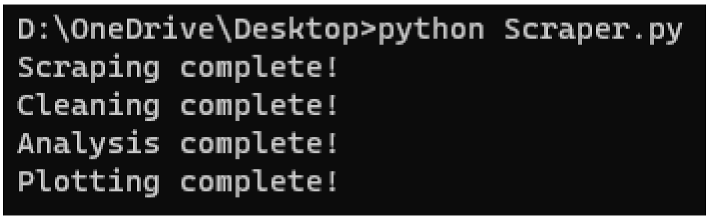
Figure 5 gives an idea of that what the processed data looks like. As can be seen in the image, there are no more issues found within the collected data and it is ready for analysis. At this stage, the program takes the processed data and performs comparisons and sentiment analysis. After producing a score for each article over a similarity threshold, it stores the final data in a new .csv with the values, as observed in Figure 6.



Source1	Article1	Objectivity1	Polarity1	Source2	Article2	Objectivity2	Polarity2	Similarity
https://www.manches	Metropolitan police ma	0.393	0.018	https://www.bbc.co.uk	Police made tough calls	0.439352724	0.037967623	0.111731844
https://www.manches	Olly Murs stuns viewer	0.41953125	0.219788707	https://www.bbc.co.uk	Coronation: Royals set	0.45372807	0.079459064	0.104925054
https://www.manches	Olly Murs stuns viewer	0.41953125	0.219788707	https://www.bbc.co.uk	King and Queen say tha	0.453726852	0.093494844	0.121818182
https://www.manches	Princess Eugenie appla	0.461784512	0.213636364	https://www.bbc.co.uk	King and Queen say tha	0.453726852	0.093494844	0.101265823
https://www.manches	Culture Secretary blam	0.426553672	0.079849341	https://www.bbc.co.uk	Police made tough calls	0.439352724	0.037967623	0.104
https://www.manches	What the Royals said or	0.455246413	0.143805365	https://www.bbc.co.uk	King and Queen say tha	0.453726852	0.093494844	0.113782051
https://www.manches	Why was Katy Perry at t	0.271825397	0.129761905	https://www.bbc.co.uk	Coronation: Royals set	0.45372807	0.079459064	0.100271003
https://www.manches	Why was Katy Perry at t	0.271825397	0.129761905	https://www.bbc.co.uk	In pictures: Coronation	0.319091711	0.097478756	0.106060606
https://www.manches	Zoe Ball replaced at las	0.332103175	0.16228355	https://www.bbc.co.uk	King and Queen say tha	0.453726852	0.093494844	0.106086957
https://www.manches	The awkward moments	0.421890866	0.063104591	https://www.bbc.co.uk	King and Queen say tha	0.453726852	0.093494844	0.101449275
https://www.manches	Three arrested in murd	0.468803419	0.061402486	https://www.bbc.co.uk	Teen victim named in B	0.554917184	0.027786404	0.135977337
https://www.manches	Prince Harry catches Bri	0.448417098	0.173866716	https://www.bbc.co.uk	King and Queen say tha	0.453726852	0.093494844	0.101503759
https://www.manches	What time is the Coron	0.413346124	0.165995999	https://www.bbc.co.uk	Coronation: Royals set	0.45372807	0.079459064	0.104408353
https://www.manches	What time is the Coron	0.413346124	0.165995999	https://www.bbc.co.uk	King and Queen say tha	0.453726852	0.093494844	0.107485605
https://www.manches	Met Police 'understand	0.40234127	0.017698413	https://www.bbc.co.uk	Police made tough calls	0.439352724	0.037967623	0.110507246

Figure 6. A representation of the final data with the similarity, polarity and subjectivity scores

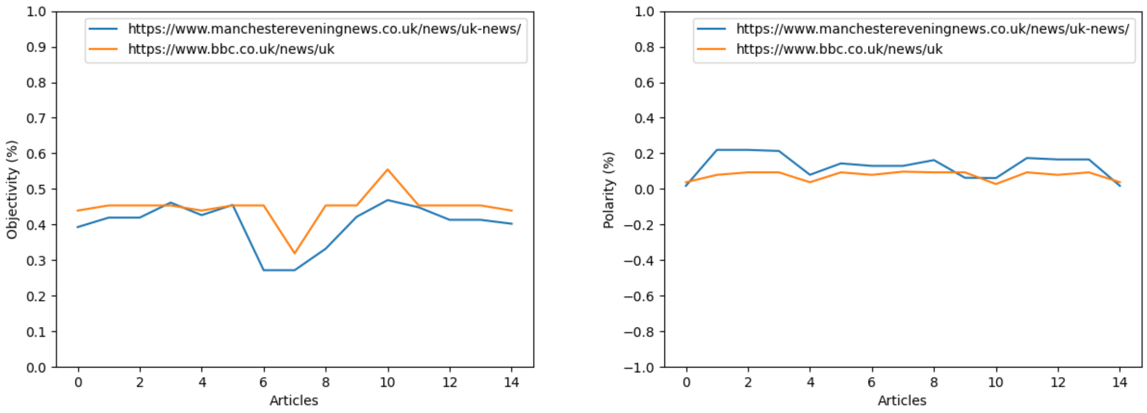
After obtaining the scores for each article, success is noted in the console, as seen in Figure 7.



```
D:\OneDrive\Desktop>python Scraper.py
Scraping complete!
Cleaning complete!
Analysis complete!
Plotting complete!
```

Figure 7. Proof the program reaches completion

Finally, the plots are produced and the scores are observed visually, as shown in Figure 8 and Figure 9.

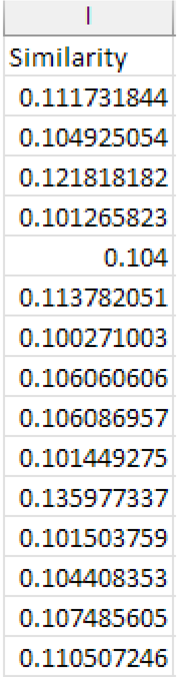


Figures 8 and 9. The final results output by the program demonstrates the subjectivity and polarity scores.

Analysis of Results

Looking at the results of the autonomous program, it currently answers each question positively but to differing degrees. It does successfully scrape news outlets for articles and stores the relevant information in an appropriate, usable manner. However, this has only been answered for two news sources with a maximum of 35 articles from the same day. There are also some notable characters which do not belong in the scraped texts before cleaning. Whilst this is likely addressed during the tokenisation of the words, it is not fully resolved either before saving the data or handled in a way that makes it fully usable for the program which could affect the final scores. Based on these results, this question is demonstrably answered positively but in a very limited use case that would need to be extended before deploying a usable application.

Moving to the next question, the program does appear to clean the data appropriately. Whilst it still fails to deal with special characters in the texts, it has successfully removed the articles which have failed to be scraped properly, leaving only the data which is most useful for the analysis. Further to this, the program does successfully make use of natural language processing. It is a light implementation and just enough to appropriately obtain the necessary information without being too complex or slowing the program down. The goal is to provide a polarity and subjectivity analysis which it performs successfully, but this is something that needs to be observed on a larger scale with a better-defined variety of articles with obviously varying scores to fully measure its accuracy. Perhaps there could be an argument or theory in the media field that does not allow for strictly subjective writing in big headlines, but this requires further research outside the scope of the project. The program gives scores for articles which only have above a similarity threshold. A drawback of the current solution is that the threshold had to be set to 10%. Ideally, this would be over 75% or at the very least 51% to ensure a chance of the article topics legitimately matching. However, in this instance, Figure 10 shows that all of the analysed data is above the required 10%.



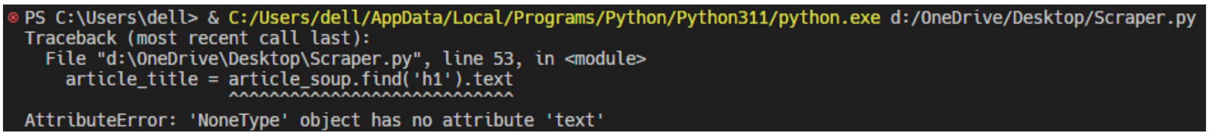
I
Similarity
0.111731844
0.104925054
0.121818182
0.101265823
0.104
0.113782051
0.100271003
0.106060606
0.106086957
0.101449275
0.135977337
0.101503759
0.104408353
0.107485605
0.110507246

Figure 10. The similarity scores generated for the analyses and plotted data

The final question to address is whether the results are expected or indicate that things could improve. With any project, there will always be something that could work better but in this instance, the lack of scalability and limited results appears to suppress the potential of the program quite a lot. This point will be reflected upon in the final section but it is worth mentioning here that the system does technically work. With some better expertise and more time with the language, the power of the system, and therefore the apparent usefulness for it, is likely to change drastically.

Project Challenges

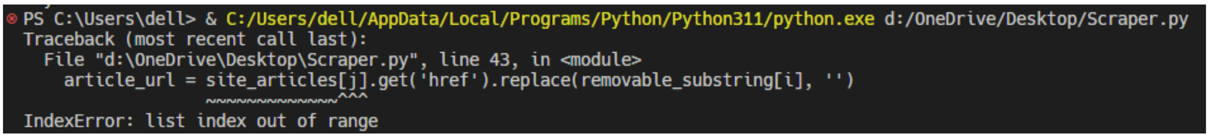
As mentioned prior, there were more than several challenges during this project. Some workarounds worked whilst other more critical errors changed the course of the project quite substantially. One of the most notable difficulties was finding a way to access different webpage elements on different websites. Given the different HTML structures of each source, the best workaround with the allocated time was to hard-code and loop through the relevant classes on each page by manually sifting through and finding what worked. This severely limits the ability of the agent to scan new websites, or one chosen by a user later in the project pipeline. Figure 9 provides an example of one of the errors seen related to finding articles objects in certain sources which had to be removed in the interest of moving forward. This could potentially be overcome using the BeautifulSoup module but more reading is required in this area.



```
PS C:\Users\dell> & C:/Users/dell/AppData/Local/Programs/Python/Python311/python.exe d:/OneDrive/Desktop/Scraper.py
  Traceback (most recent call last):
    File "d:\OneDrive\Desktop\Scraper.py", line 53, in <module>
      article_title = article_soup.find('h1').text
                      ^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  AttributeError: 'NoneType' object has no attribute 'text'
```

Figure 11. An error caused by failure to acquire article objects

The second most limiting challenge was finding a way to retain the most data, especially taking into account the previous issue. Attempting to scrape more than 35 articles from each website proved fruitless and resulted in the error seen in Figure 12.



```
⊗ PS C:\Users\dell> & C:/Users/dell/AppData/Local/Programs/Python/Python311/python.exe d:/OneDrive/Desktop/Scraper.py
  Traceback (most recent call last):
    File "d:\OneDrive\Desktop\Scraper.py", line 43, in <module>
      article_url = site_articles[j].get('href').replace(removable_substring[i], '')
                    ~~~~~~~~~~~~~~~~^^^
  IndexError: list index out of range
```

Figure 12. An error caused by reaching the end of the articles in the list

Even with some sources less than 35 produced this error so unfortunately these were also dropped from the project in the interest of progress. These two issues absolutely must be addressed in future work if the agent is to work as fully intended. The consequences of these made it difficult to retain articles with high similarity since the analysed data was extremely limited. The hope is that alleviating the above would do the same for improving the similarity between articles. More minor issues include unknown characters in the clean texts. The impacts of this are not fully understood but are worth investigating after measuring the effects of the more major issues in the project. Similarly, removing the article text that relates to newsletters or the website etc. needs to be removed for each source. This could likely be resolved by including tokenising these and adding them to the stopwords list. The final issue relates to how the similarity score is calculated. The Jaccard index only takes into account the length of sets rather than the semantics behind the words included. Word embeddings would likely be a better approach and should be investigated moving forward.

Conclusion

A large portion of this project was successful. The code written demonstrates the desired outcomes and provides a well-defined pipeline moving forward. A useful intelligent agent has been created but the autonomous aspects need work. If the agent stayed connected to the sites and continually produced results via sentiment analysis then the requirements of the project would have been fully met. Unfortunately, this now falls under future work but that completed up-to-this point proves the potential and direction the project can take. The previously mentioned limitations have held the project back by some substantial margin but gaining some of the desired results was still possible even with a lot of difficulty and a lack of personal knowledge in Python and its modules. Considering the future work angle, the project would best be finalised with a sleek graphical user interface that revolves around an interactive slider on a subjectivity-polarity map. This would allow the user to select their desired scores for the program to produce the closest matching news source. Another useful piece of information for the user would be using a news outlet's historical data to measure the subjectivity-polarity scores over some time until the present day. This could also lead to another small machine learning implementation which predicts the direction of the publication in this format. Finally, before moving forward it could prove fruitful to perform a small analysis of the media industry and determine whether there are generally accepted levels of subjectivity and polarity in writing across the board and what difference there is between different types of outlets etc.

References

Agents in Artificial Intelligence (no date). Available at: https://www.javatpoint.com/agents-in-ai (Accessed: 9 May 2023).

Alam, G. *et al.* (2022) 'Applications of artificial intelligence in water treatment for optimization and automation of adsorption processes: Recent advances and prospects', *Chemical Engineering Journal*, 427, p. 130011. Available at: https://doi.org/10.1016/J.CEJ.2021.130011.

Amos, D. (2022) *A Practical Introduction to Web Scraping in Python*. Available at: https://realpython.com/python-web-scraping-practical-introduction/ (Accessed: 9 May 2023).

API reference — pandas 2.0.1 documentation (no date). Available at: https://pandas.pydata.org/docs/reference/index.html#api (Accessed: 8 May 2023).

Baghbani, A. *et al.* (2022) 'Application of artificial intelligence in geotechnical engineering: A state-of-the-art review', *Earth-Science Reviews*, 228, p. 103991. Available at: https://doi.org/10.1016/J.EARSCIREV.2022.103991.

Bhutani, B. *et al.* (2019) 'Fake News Detection Using Sentiment Analysis', *2019 12th International Conference on Contemporary Computing, IC3 2019* [Preprint]. Available at: https://doi.org/10.1109/IC3.2019.8844880.

Bird, S., Klein, E. and Loper, E. (2009) *Natural Language Processing with Python*, *Journal of Endodontics*. O'Reilly. Available at: https://www.oreilly.com/library/view/natural-language-processing/9780596803346/ (Accessed: 8 May 2023).

Brown, S. (2021) *Machine learning, explained*. Available at: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained (Accessed: 25 April 2023).

Carrillo-Perez, F. *et al.* (2022) 'Applications of artificial intelligence in dentistry: A comprehensive review', *Journal of Esthetic and Restorative Dentistry*, 34(1), pp. 259–280. Available at: https://doi.org/10.1111/JERD.12844.

Choudhuri, A. (2023) *Types of intelligent agent*. Available at: https://www.probegroup.com.au/blog/types-of-intelligent-agent (Accessed: 8 May 2023).

dataframe_image (2023). Available at: https://pypi.org/project/dataframe-image/ (Accessed: 8 May 2023).

Facer, C. (2018) *How to Calculate Jaccard Coefficients in Displayr Using R*. Available at: https://www.r-bloggers.com/2018/10/how-to-calculate-jaccard-coefficients-in-displayr-using-r/ (Accessed: 9 May 2023).

Ganesan, K. (no date) *What are Stop Words?* Available at: https://kavita-ganesan.com/what-are-stop-words/ (Accessed: 9 May 2023).

Geitgey, A. (2018) *Natural Language Processing is Fun!* Available at: https://medium.com/@ageitgey/natural-language-processing-is-fun-9a0bff37854e (Accessed: 25 April 2023).

Gupta, S. (2018) *Sentiment Analysis: Concept, Analysis and Applications*. Available at: https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17 (Accessed: 9 May 2023).

Jang, W.D. *et al.* (2022) ‘Applications of artificial intelligence to enzyme and pathway design for metabolic engineering’, *Current Opinion in Biotechnology*, 73, pp. 101–107. Available at: https://doi.org/10.1016/J.COPBIO.2021.07.024.

Karabiber, F. (no date) *Jaccard Similarity — LearnDataSci*. Available at: https://www.learndatasci.com/glossary/jaccard-similarity/ (Accessed: 8 May 2023).

Loria, S. (no date) *TextBlob: Simplified Text Processing*. Available at: https://textblob.readthedocs.io/en/dev/ (Accessed: 8 May 2023).

Majid, A. (2023) *Most popular websites for news in the UK: Monthly top 50 listing*. Available at: https://pressgazette.co.uk/media-audience-and-business-data/media_metrics/most-popular-websites-news-uk-monthly-2/ (Accessed: 8 May 2023).

Manoharan, Dr.S. (2019) ‘AN IMPROVED SAFETY ALGORITHM FOR ARTIFICIAL INTELLIGENCE ENABLED PROCESSORS IN SELF DRIVING CARS’, *Journal of Artificial Intelligence and Capsule Networks*, 2019(2), pp. 95–104. Available at: https://doi.org/10.36548/jaicn.2019.2.005.

matplotlib.pyplot — Matplotlib 3.7.1 documentation (no date). Available at: https://matplotlib.org/stable/api/pyplot_summary.html#module-matplotlib.pyplot (Accessed: 8 May 2023).

Menzli, A. (2023) *Tokenization in NLP: Types, Challenges, Examples, Tools*. Available at: https://neptune.ai/blog/tokenization-in-nlp (Accessed: 9 May 2023).

Neogi, A.S. *et al.* (2021) ‘Sentiment analysis and classification of Indian farmers’ protest using twitter data’, *International Journal of Information Management Data Insights*, 1(2), p. 100019. Available at: https://doi.org/10.1016/J.JJIMEI.2021.100019.

NumPy Reference — NumPy v1.24 Manual (2022). Available at: https://numpy.org/doc/stable/reference/index.html#reference (Accessed: 8 May 2023).

Pascual, F. (2022) *Getting Started with Sentiment Analysis using Python*. Available at: https://huggingface.co/blog/sentiment-analysis-python (Accessed: 9 May 2023).

requests (no date). Available at: https://pypi.org/project/requests/ (Accessed: 8 May 2023).

Richardson, L. (no date) *Beautiful Soup Documentation —Beautiful Soup 4.12.0 documentation*. Available at: https://www.crummy.com/software/BeautifulSoup/bs4/doc/# (Accessed: 8 May 2023).

Taj, S., Shaikh, B.B. and Fatemah Meghji, A. (2019) ‘Sentiment analysis of news articles: A lexicon based approach’, *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies, iCoMET 2019* [Preprint]. Available at: https://doi.org/10.1109/ICOMET.2019.8673428.

Try and Except in Python (no date). Available at: https://pythonbasics.org/try-except/ (Accessed: 8 May 2023).

Zhang, C. and Lu, Y. (2021) ‘Study on artificial intelligence: The state of the art and future prospects’, *Journal of Industrial Information Integration*, 23, p. 100224. Available at: https://doi.org/10.1016/J.JII.2021.100224.

Zhang, T. *et al.* (2022) ‘Natural language processing applied to mental illness detection: a narrative review’, *npj Digital Medicine 2022 5:1*, 5(1), pp. 1–13. Available at: https://doi.org/10.1038/s41746-022-00589-7.