Political Bias Search Engine

Written Report

Group Members	Kieran Flynn ➔ 16334663 ➔ kieran.flynn36@mail.dcu.ie Kevin McGonigle ➔ 16318486 ➔ kevin.mcgonigle2@mail.dcu.ie James Miles ➔ 16349533 ➔ james.miles2@mail.dcu.ie Eoin O'Brien ➔ 15324971 ➔ eoin.obrien82@mail.dcu.ie
Programme	Computer Applications & Software Engineering Year 4 (CASE4)
Module	Search Technologies (CA4009)
Date of Submission	17/12/2019
Word Count (excluding disclaimer, table of contents, abstract & references)	3008

Plagiarism Disclaimer

A report submitted to Dublin City University, School of Computing for module CA4009: SearchTechnologies, 2019/2020.

I understand that the University regards breaches of academic integrity and plagiarism as grave and serious.I have read and understood the DCU Academic Integrity and Plagiarism Policy. I accept the penalties that may be imposed should I engage in practice or practices that breach this policy.

I have identified and included the source of all facts, ideas, opinions, viewpoints of others in the as-signment references. Direct quotations, paraphrasing, discussion of ideas from books, journal articles,internet sources, module text, or any other source whatsoever are acknowledged and the sources cite dare identified in the assignment references.

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.By signing this form or by submitting this material online I confirm that this assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this orany other course of study.

By signing this form or by submitting material for assessment online I confirm that I have read and understood DCU Academic Integrity and Plagiarism Policy (available at: http://www.dcu.ie/registry/examinations/index.shtml)

Name(s): Kieran Flynn, Kevin McGonigle, James Miles, Eoin O'Brien

Date: 17/12/2019

0. Table of Contents

1. Abstract

This report will describe the background, motivation, design and evaluation plan for a new political information retrieval system, with the goal of providing the user with factual and unbiased information that satisfies any information need they may have relating to politics.

The proposed search engine will seek to mitigate the effects of the recent growth in "fake news" publications and political bias present in media by accounting for veracity and neutrality when ranking relevant documents. To accomplish this, the search engine will initially utilise the Bing Web Search API[1] to perform the initial retrieval of web documents that are relevant to the user's query.

From there, our system will parse the contents of the returned documents and classify the sentiment, bias and accuracy of information present in each. These classifications will be used to as part of a ranking algorithm to determine the order in which the documents are displayed to the end-user, and will be stored as metadata for optimisation of future requests.

The classification will be informed by recursive learning from user feedback and the search engine will be evaluated using a collection of human-evaluated internal test documents, information needs and corresponding queries.

2. Introduction

2.1 Background

In recent years, the world has seen an increase in the presence of partisan media bias with respect to politics (Prior, 2013)[2] along with the popularisation of "fake news" stories surrounding political matters. Examples of both are glaring amidst major political events such as the 2016 United States presidential election, the Brexit referendum of 2016 and the recent 2019 UK general election, with negative reporting and fake news at the forefront of media coverage throughout.

As a result, it has become increasingly difficult to source reliable and impartial information, particularly online, with 58% and 41% of individuals that were surveyed claiming a drop in their level of trust in stories on social media and online-only news outlets respectively (Cooke, 2017)[3].

---

[1] Microsoft. "What Is the Bing Web Search API?" *Azure Cognitive Services*, Microsoft, 5 Dec. 2019, docs.microsoft.com/en-gb/azure/cognitive-services/bing-web-search/overview
[2] Prior, Markus. "Media and political polarization." *Annual Review of Political Science*, vol.16, May 2013: 101-127. Annual Reviews, doi:10.1146/annurev-polisci-100711-135242
[3] Cooke, Kirsty. "'Fake news' reinforces trust in mainstream news brands" *Kantar*, 31 Oct. 2017, https://uk.kantar.com/business/brands/2017/trust-in-news.

Fueling this growing distrust are recent revelations of scandals relating to personal data being used to influence political campaign strategies. Perhaps the most notable example being the Facebook - Cambridge Analytica (CA) scandal in which the personal information of ≈87 million individuals were improperly provided to CA[4] and used to influence the 2016 presidential election[5,6].

The result is a public with faltering trust in media and politicians alike that lack a method of obtaining unbiased, factual information about presidential candidates, political parties and referendum topics; leaving their vote in jeopardy of being swayed by biased, false or potentially unfairly targeted information.

2.2 Motivation

The motivation for our desire to design and develop a political information retrieval system stems from the overwhelming and united demand for truthful and impartial news coverage of political events (Mitchell et al, 2018)[7].

Partisan bias in the media can have a substantial influence over the consumer's political attitudes, ideals, and even choice of vote (McCombs, 2005)[8]. Therefore, it is our view that providing a method for searching and browsing unbiased and accurate information is essential for maintaining a fair environment in which the consumer can form their individual political ideals without the burden of bias or dishonesty.

The lack of a comprehensive tool that is specifically designed to provide reputable and non-partisan political information to the user leaves a gap in the marketplace to be filled by the proposed system. We view the prospective uniqueness and scope for expansion of this proposed search engine as major motivating factors, on top of the existing demand.

---

[4] Schroepfer, Mike. "An Update on Our Plans to Restrict Data Access on Facebook." Facebook, 4 Apr. 2018, https://about.fb.com/news/2018/04/restricting-data-access.

[5] Davies, Harry. "Ted Cruz using firm that harvested data on millions of unwitting Facebook users." The Guardian, 11 Dec. 2015, https://www.theguardian.com/us-news/2015/dec/11/senator-ted-cruz-president-campaign-facebook-user-data.

[6] Cheshire, Tom. "Behind the Scenes at Donald Trump's UK Digital War Room." Sky News, Sky, 22 Oct. 2016, news.sky.com/story/behind-the-scenes-at-donald-trumps-uk-digital-war-room-10626155.

[7] Mitchell, Amy, et al. "Publics Globally Want Unbiased News Coverage, but Are Divided on Whether Their News Media Deliver." Pew Research Center, 11 Jan 2018, https://www.pewresearch.org/global/2018/01/11/publics-globally-want-unbiased-news-coverage-but-are-divided-on-whether-their-news-media-deliver.

[8] McCombs, Maxwell. "A Look at Agenda-setting: past, present and future" *Journalism Studies*, vol. 6, no. 4, 2005, pp. 543-557. Taylor & Francis, doi:10.1080/14616700500250438

2.3 Objectives

2.3.1 Relevance

We aim to produce an information retrieval system that will successfully and consistently retrieve web-based political content that is as close to wholly relevant to the user's query, and corresponding information need, as possible.

2.3.2 Ranking & Filtering

One of our primary objectives for the proposed system is to devise and implement a ranking algorithm that accounts for the presence of bias, dishonesty and sentiment, as well as considering the recency of publication of the returned documents and the relevance rank that was initially obtained.

We also aim to allow the user of the system to configure how results should be ranked, as well as providing a method for filtering the returned collection based on the prior mentioned attributes.

2.3.3 Learning from Feedback

We intend to have our implemented system improve the classification of bias, veracity and relevance by learning from both supplied and observed user feedback. We aim to provide the user with a means for directly supplying feedback on each of these attributes for any retrieved document that they navigate to.

On top of a direct approach, we propose the application of interactive evaluation that involves deriving conclusions about document relevance and information need satisfaction by observing a number of variables such as number of queries entered, number of documents opened and time spent searching (Kelly, 2009)[9].

2.3.4 Incentivise Change

One of our external goals for our system is for it to bring about an end, or at least an improvement, to the present culture of falsified and biased reporting with the sole purpose of generating web page visits. We intend to do this inherently through rewarding impartial and factual journalism, in the hope that it will incentivise news outlets and politicians alike to stray from the existing paradigm of opposition degradation, and exaggeration or fabrication of political stories.

---

[9] Kelly, Diane. "Methods for Evaluating Interactive Information Retrieval Systems with Users" *Foundations and Trends in Information Retrieval*, vol. 3, no. 1&2, 28 Apr. 2009, pp. 1–224. Now Publishers, doi:10.1561/1500000012

3. User Analysis

3.1 User Characteristics

Understanding the characteristics of users who will interact with a system is a key task for any system design. In our case, a typical user is hard to define as politics is a diverse and broad topic and as such our system will need to account for a massive variety in user characteristics. Our users will likely range in technological ability and political knowledge, therefore the user interface must be intuitive to use and easy to understand.

3.2 Operational Scenarios

In this section, we examine a typical operation scenario for users of differing levels of political knowledge and familiarity.

3.2.1 Weak Political Knowledge

Jenna is a third level student who understands the worrying confusion between entertainment and news. She often finds herself cross-checking news article's facts and sources, with her research often revealing discrepancies, or finding them to be misleading, biased or false. Jane is looking for a consistent reliable source of information that values unbiased factual articles rather than those whose objective is purely to generate hits through controversial and misleading headlines.

3.2.2 Moderate Political Knowledge

James is a middle-aged father whose son has recently shown him the effect that Cambridge Analytica had over the voting population of the United States. Looking to further his political knowledge before the upcoming election, he has begun looking for information from reliable sources without a political, financial or other agenda. He wishes there was a single source that could differentiate between the real and the fake, removing false news and webpages that produce or distribute misleading content.

3.2.3 Strong Political Knowledge

Joe is a campaign director for candidates in the upcoming Democratic National Convention that will decide who competes against current American president Donald Trump for the 2020 presidency. As part of his job, he needs to be aware of any news regarding his client and the candidates that his client is competing against. With an abundance of false narratives being displayed about his client, their competition and the topics they are questioned about, Joe needs to be able to provide his client with accurate information to avoid backlash from the media and help his clients get the votes needed to win. Joe feels that his job would be much easier if there was a tool he could use to separate the truth from these false narratives online and give his client the facilities they need to triumph in their campaign over the competition.

Fig 3.2.4 Typical Users

3.3 Typical User Queries

There are a number of common characteristics that we would expect to see when it comes to user search queries within our proposed system. These characteristics are seen below.

3.3.1 Short Queries

A typical search query is approximately 2.5 words in length (Spink et al, 2001)[10]. As such, we will need to ensure that our system can accurately identify the key information need(s) of the user. In cases where the need is ambiguous, the system can return documents that vary in type as well as interpretations of the query. This allows users themselves to find documents that satisfy their need.

3.3.2 Uninformed Queries

The nature of Information Retrieval is that there is something the user does not know. As such, users are often unable to describe their information need effectively (anomalous state of knowledge (ASK) problem). After an initial attempt to address the information need our system can provide documents which may help them form a more effective reformulated query at which point they can form another search.

3.3.3 Broad Range of Queries

A user's information need may be from a massive range of topics, even when restricted to the predefined domain of politics. Our system will need to be able to

[10] Spink, Amanda, et al. "Searching the Web: The Public and Their Queries" *Journal of the American Society for Information Science and Technology*, vol. 52, no. 3, 25 Jan. 2001, pp. 226-234. Wiley Online Library, doi:10.1002/1097-4571(2000)9999:9999<::AID-ASI1591>3.0.CO;2-R

handle variance in the need of the user and return documents that can effectively satisfy said need.

3.3.4 Biased Queries

The majority of users will have their own intrinsic opinions and beliefs. As such, their queries for information needs may be biased. The system will have to be able to identify these biases and account for them when performing searches.

4. Scientific Functional Description

4.1 Algorithms & Technologies

4.1.1 Information Retrieval - Bing Web Search API

For the initial information retrieval, we propose the use of Microsoft Azure Cognitive Services' Bing Web Search API. This RESTful API will allow the system to search the web for relevant information using the Bing search engine which provides a number of useful functionalities that may be used to enhance our own application such as related queries, spelling suggestions and custom ranking and filtering[11].

We opted for the use of Bing over Google's equivalent custom search product due to the limitations on search domains placed by Google upon reaching a query quota. With Bing, we are free to search the entire web irrespective of the amount of queries submitted.

4.1.2 Bias Classification

For determining the magnitude and direction of bias within a document, we propose the adoption of a long short-term memory architecture (LSTM) (Hochreiter & Schmidhuber, 1997)[12]. LSTM is a form of recurrent neural network (RNN) that has recently shown great promise as a deep learning solution to a number of natural language processing challenges, frequently outperforming classing techniques . This is largely due to the deep artificial understanding of the structure of language, obtained by being tasked to predict the next word in a sentence and thus learning the underlying semantic structure (Misra & Basak, 2016)[13].

It has been documented that LSTM units require very large training datasets with considerable overlap. We are optimistic that the combination of collections of

[11] Microsoft. "Bing Web Search" *Azure Cognitive Services*, Microsoft, 2019, https://azure.microsoft.com/en-us/services/cognitive-services/bing-web-search-api
[12] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural Computation*, vol.9, no. 8, 15 Nov. 1997, pp. 1735-1780. The MIT Press, doi:10.1162/neco.1997.9.8.1735
[13] Misra, Arkajyoti, and Sanjib Basak. "Political Bias Analysis", *Department of Computer Science,* Stanford University, 2016, https://cs224d.stanford.edu/reports/MisraBasak.pdf

internal training datasets, external datasets such as ontheissues[14] and user feedback on retrieved relevant documents.

4.1.3 Sentiment Analysis

We intend to perform sentiment analysis to identify the polarity of sentiment contained within each retrieved relevant document; that is to say, the extent to which it could be considered to be of a positive, negative or neutral opinion. This will be accomplished through a process of feature extraction in which words (excluding stop words and with morphological variants being represented by the base form) will be vectorised based on their frequency within the document (bag-of-words model) (McTear, 2016). The vectorised text will then be used to inform a Naive Bayes Classifier that will be trained to classify polarity of sentiment using various datasets of textual input that has been independently rated for sentiment by humans (e.g. nytEditorialSnippets provided by VADER[15]).

4.1.4 Veracity Classification

We plan to implement a method of classifying the likely truthfulness and reputability of the contents of retrieved documents. This will be done through a combination of Naive Bayes Classification, trained using a human-verified "fake news" dataset on Kaggle, and an approach similar to that of TrustRank[16] in which a collection of web pages are evaluated by an expert to be used as a reputable seed for utilising the web's link structure to locate similarly reliable pages.

4.1.5 Ranking

Our algorithm looks to place an equal importance on a document's relevance, neutrality with respect to both bias and sentiment, and accuracy. As such, we have devised the following formula for generating a rank score, by which documents will be ranked in descending order by default.

[14] OnTheIssues, "OnTheIssues.org - Candidates on the Issues", *OnTheIssues*, 2019, https://archive.ontheissues.org/default.htm
[15] Hutto, C.J., and Eric Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text" VADER, 2014, https://github.com/cjhutto/vaderSentiment
[16] Gyöngyi, Z., et al. "Combating Web Spam with TrustRank" *Proceedings of the Thirtieth international conference on Very large data bases*, vol. 30, no. 1, 31 Aug. 2004, pp. 576-587. VLDB Endowment, doi:10.1016/B978-012088469-8/50052-8

$$RS = \frac{V(1-B)(1-|S|)}{RR}$$

Where:

RS = Rank Score ($0 \le RS \le 1$)
RR= Relevance Rank ($1 \le RR$)
B = Bias Classification ($0 \le B \le 1$)
S = Sentiment Classification ($-1 \le S \le 1$)
V = Veracity Classification ($0 \le V \le 1$)

This formula assumes the normalisation of B, S and V to fall within the ranges specified. The relevance rank in question is the order in which a given document was returned from the Bing Web Search API (i.e. the topmost document will have an RR of 1, the second topmost will have an RR of 2 etc.). Choosing relevance rank as the denominator will produce the effect of enforcing a higher weight for relevance for documents that were near the top of the returned collection, and evening out towards the bottom of the collection where the other metrics will have more of an effect on the rank score.

4.1.6 Iterative Recursive Learning

Our system will make use of feedback provided by the user after searching. This feedback aims to gauge the validity of the returned documents in 3 areas:

- Relevance
- Bias
- Sentiment

The user will be able to inform the system of any mistakes or inaccuracies in the above areas. By making use of feedback, we can recursively train our system to be more accurate and therefore fulfill the user's information need more reliably.

4.2 Overall Architecture

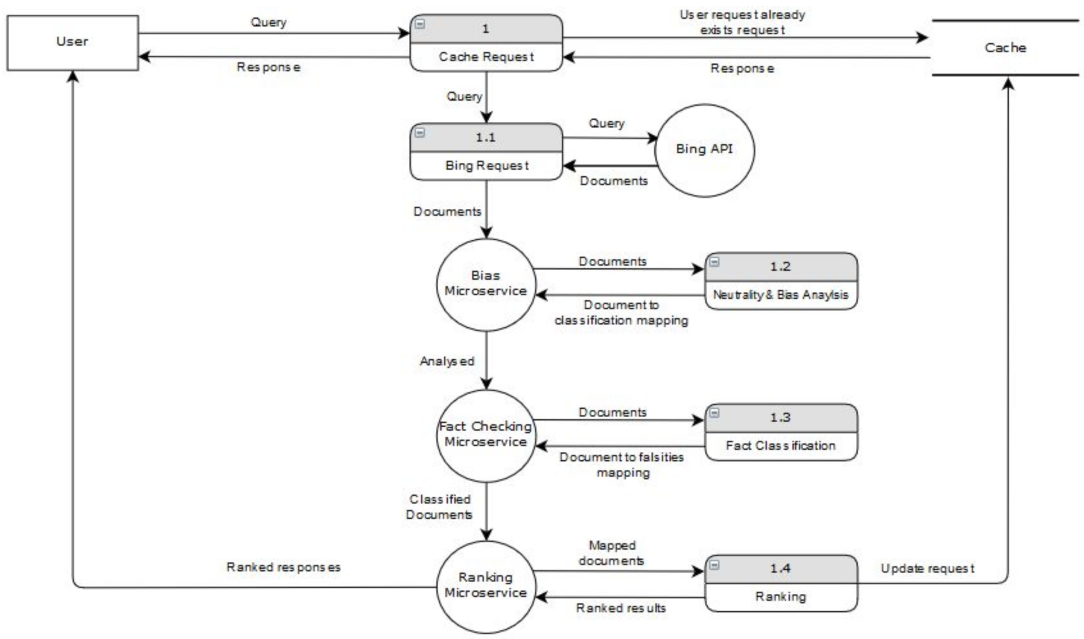
4.2.1 Data Flow

The general flow of data through the proposed system is to be implemented as follows:

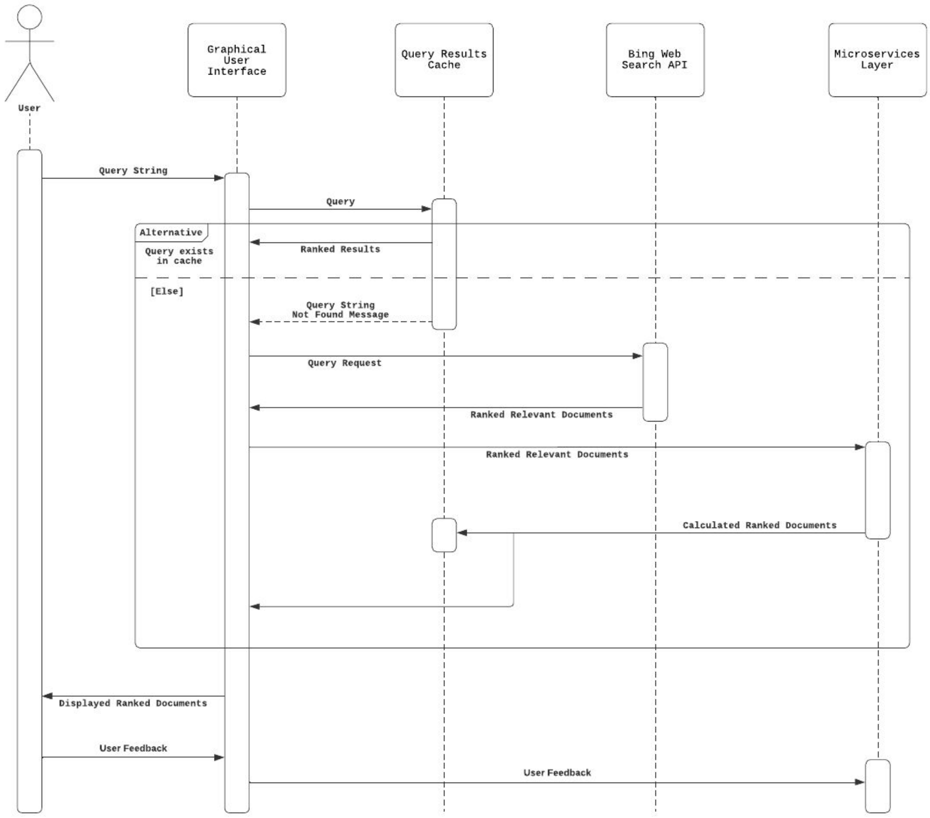
1. The user loads the graphical user interface on their device and enters a search query.
2. The system checks if the entered query has previously been entered and exists in the cache.
    - If yes, the previously obtained results are displayed on the GUI. The flow of data terminates.

3. If no, the system will use the query to request ranked relevant documents from the Bing Web Search API, which will be returned in a JSON response object.
4. The system will send the relevant documents list to the microservices layer for the computation of bias and factuality judgements, and for re-ranking of the list accordingly.
5. The system displays the ranked list to the user, and logs it in the cache for optimisation and consistency.
6. The system receives user feedback on relevancy, bias and factuality and adjusts accordingly for future usages.

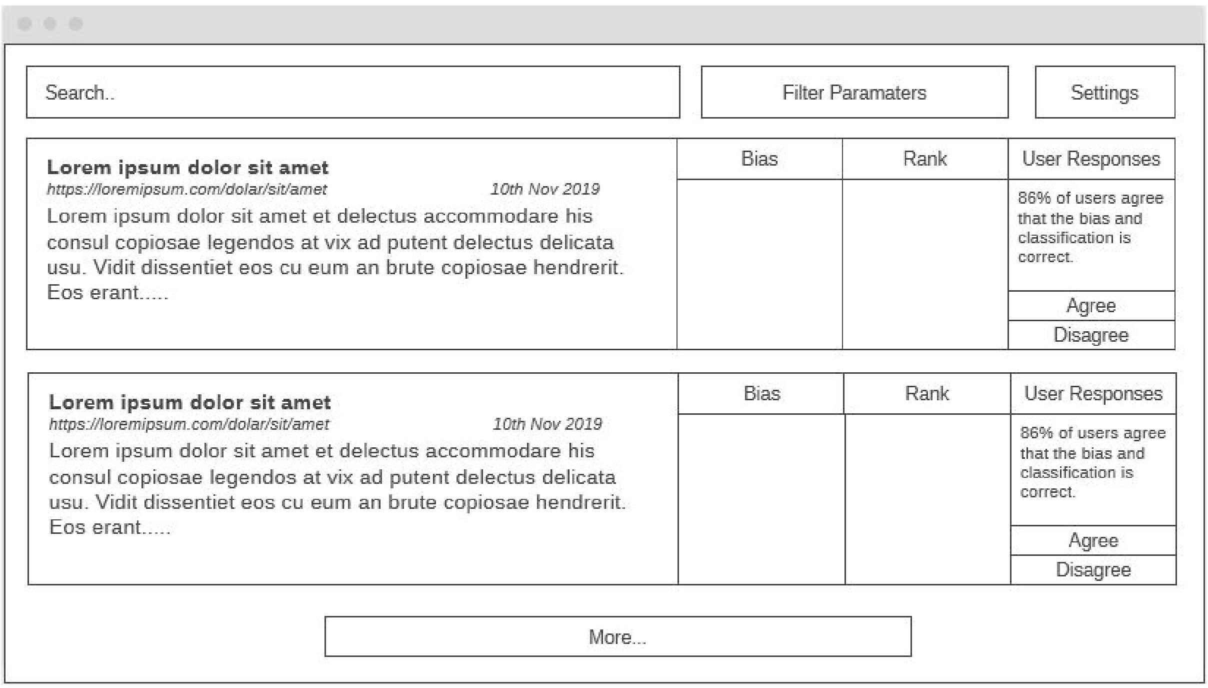
4.2.1.1 Data Flow Diagram



4.2.1.2 Sequence Diagram



4.2.2 User Interface Wireframe



4.3 Limitations

4.3.1 Subjectivity

The subjective nature of sentiment, intent & bias is a significant limitation with our proposed system. These factors, along with similar metrics, are based on a specific user's beliefs or opinions and therefore will likely change from user to user.

4.3.2 Feedback Reliance

The system is reliant on accurate user feedback in order to reliably return factually accurate and bias free documents where possible. This feedback is crucial because determining factors such as sentiment or bias is extremely difficult without user input.

4.4 Assumptions

4.4.1 User Information Needs

It is assumed that a user's information needs will be strictly related to politics. The scope of the project means that a predefined domain of information needs is required and therefore anything outside of this scope will not be considered by the system.

4.4.2 Storage Assumptions

It is assumed that the system will meet the necessary storage requirements to store all metadata or metrics for documents that are calculated and gathered.

5. Evaluation

5.1 Strategy

5.1.1 Internal Test Data

Our system will also make use of a large collection of internal testing documents, information needs and corresponding queries with human-verified binary results for each metric. This is crucial to our evaluation plan as without it the system would struggle to identify relevant documents accurately and ensure that all bias & sentiment analysis is actually correct. This internal test data will also be used in conjunction with the methods in the next section.

5.1.2 System Testing

We plan to carry out rigorous system testing to ensure all facets of the search engine are working as desired.

To facilitate consistent system testing, the results returned by the Bing Web Search API for a given query will be stored. Subsequently, should the same query be entered again, the stored results will be returned and the API will not be used. The reason for this is that the Bing API is an external resource over which we have no direct control. Therefore there is no guarantee that discrepancies between returned results for the same query will not arise.

5.2 Objectives

5.2.1 Sentiment Analysis & Bias Classification Accuracy

Our system will aim for approximately 80% agreement with user feedback & gold standards (Correlating with human-agreement ratio).[17]

5.2.2 Data & Metrics

We will aim to maximise precision, recall and mean average precision (MAP) within our system. While it can be difficult to manage the trade offs between maximising precision and recall respectively, we can strive to find the right balance between reducing the number of false positives and negatives recovered by the system[18]. By maximising these various metrics we hope that our system will be accurate, reliable and useful to all users who interact with it.

6. Conclusion

6.1 Summary

In conclusion, the basis of this project is utilising the latest text and data retrieval technologies to combat the modern day problem of false political information. This system will integrate relevance searching models and API's alongside machine learning algorithms for classification and sentiment analysis to provide users with a ranking of accurate, unbiased sources for the stories they wish to learn further about.

---

[17] Ogneva, M., "How Companies Can Use Sentiment Analysis to Improve Their Business." *Mashable*, 19 Apr. 2010, https://mashable.com/2010/04/19/sentiment-analysis

[18] Manning, Christopher D., et al. *Introduction to Information Retrieval*. Cambridge University Press, 2008

6.2 Next Steps

To pursue this project in the future, there are a number of key areas that we would need to explore. In order for the system to incorporate its classification and Sentiment Analysis aspects there needs to be significant steps taken in developing a Machine Learning model to handle both features. Following on from this, there will be further data handling in terms of caching the metadata of documents retrieved from the system and creating a database for it. There will also have to be a period of testing in the development cycle to assemble our own collection of test documents, informational requirements and search queries & identify the reaction of the system to said resources.

All of this is on top of the actual deployment of the system & begin iteratively taking in feedback from users and improving upon the presentation and performance of the system over time.

7. References

Microsoft. "What Is the Bing Web Search API?" Azure Cognitive Services, Microsoft, 5 Dec. 2019, docs.microsoft.com/en-gb/azure/cognitive-services/bing-web-search/overview

Prior, Markus. "Media and political polarization." Annual Review of Political Science, vol.16, May 2013: 101-127. Annual Reviews, doi:10.1146/annurev-polisci-100711-135242

Cooke, Kirsty. "'Fake news' reinforces trust in mainstream news brands" Kantar, 31 Oct. 2017, https://uk.kantar.com/business/brands/2017/trust-in-news

Schroepfer, Mike. "An Update on Our Plans to Restrict Data Access on Facebook." Facebook, 4 Apr. 2018, https://about.fb.com/news/2018/04/restricting-data-access.

Davies, Harry. "Ted Cruz using firm that harvested data on millions of unwitting Facebook users." The Guardian, 11 Dec. 2015, https://www.theguardian.com/us-news/2015/dec/11/senator-ted-cruz-president-campaign-facebook-user-data.

Cheshire, Tom. "Behind the Scenes at Donald Trump's UK Digital War Room." Sky News, Sky, 22 Oct. 2016, news.sky.com/story/behind-the-scenes-at-donald-trumps-uk-digital-war-room-10626155

Mitchell, Amy, et al. "Publics Globally Want Unbiased News Coverage, but Are Divided on Whether Their News Media Deliver." Pew Research Center, 11 Jan 2018, https://www.pewresearch.org/global/2018/01/11/publics-globally-want-unbiased-news-coverage-but-are-divided-on-whether-their-news-media-deliver

McCombs, Maxwell. “A Look at Agenda-setting: past, present and future” Journalism Studies, vol. 6, no. 4, 2005, pp. 543-557. Taylor & Francis, doi:10.1080/14616700500250438

Kelly, Diane. “Methods for Evaluating Interactive Information Retrieval Systems with Users” Foundations and Trends in Information Retrieval, vol. 3, no. 1&2, 28 Apr. 2009, pp. 1–224. Now Publishers, doi:10.1561/1500000012

Spink, Amanda, et al. “Searching the Web: The Public and Their Queries” Journal of the American Society for Information Science and Technology, vol. 52, no. 3, 25 Jan. 2001, pp. 226-234. Wiley Online Library, doi:10.1002/1097-4571(2000)9999:9999<::AID-ASI1591>3.0.CO;2-R

Microsoft. “Bing Web Search” Azure Cognitive Services, Microsoft, 2019, https://azure.microsoft.com/en-us/services/cognitive-services/bing-web-search-api

Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural Computation, vol.9, no. 8, 15 Nov. 1997, pp. 1735-1780. The MIT Press, doi:10.1162/neco.1997.9.8.1735

Misra, Arkajyoti, and Sanjib Basak. “Political Bias Analysis”, Department of Computer Science, Stanford University, 2016, https://cs224d.stanford.edu/reports/MisraBasak.pdf

OnTheIssues, “OnTheIssues.org - Candidates on the Issues”, OnTheIssues, 2019, https://archive.ontheissues.org/default.htm

Hutto, C.J., and Eric Gilbert, “VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text” VADER, 2014, https://github.com/cjhutto/vaderSentiment

Gyöngyi, Z., et al. “Combating Web Spam with TrustRank” Proceedings of the Thirtieth international conference on Very large data bases, vol. 30, no. 1, 31 Aug. 2004, pp. 576-587. VLDB Endowment, doi:10.1016/B978-012088469-8/50052-8

Ogneva, M., “How Companies Can Use Sentiment Analysis to Improve Their Business.” Mashable, 19 Apr. 2010, https://mashable.com/2010/04/19/sentiment-analysis

Manning, Christopher D., et al. Introduction to Information Retrieval. Cambridge University Press, 2008