Social Media Bias Analyzer

Progress report

Joshua Fitzmaurice
Department of Computer Science
University of Warwick

1 Introduction

Bias is the overrepresentation of a particular topics in an individuals social media feed. This bias could be representative of the individuals interests, or it could be representative of the interests of the social media platform. The latter is where the notion of bias becomes complicated. Is my feed bias if it shows me only posts on topic A, if the only posts on the social media platform are of topic A? For the purpose of this report, we will take the stand that bias refers to the deviation from the "norm".
Bias Definition: the deviation of a users social media feed from the distribution of available posts.

Bias in social media can easily be seen by comparing users' social media. In fact, this can be shown with ease by just taking a look at my Instagram's "Explore page", and compare it to another's.

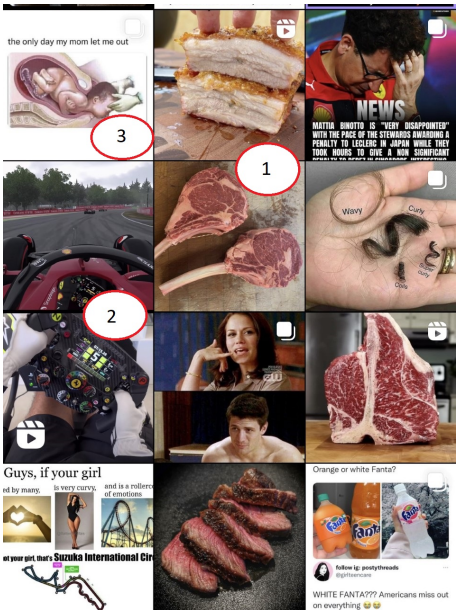

Figure 1: My Instagram for you page

Here we can notice a few common themes/biases: 1. Food, 2. Formula 1, 3. Memes. We want to be able to identify these biases for users so they can get an overview of the type of content they are receiving from social media.
With social media recommender systems programmed to entice users with content they will enjoy (Shin (2020)), it is common for similar groups of posts to be observed by a user if they have recently liked, commented, or viewed similar posts (Instagram).
Although the content in the feed only shows a small number of topics, we can not guarantee bias, as we are unaware of the distribution of all posts on the platform.

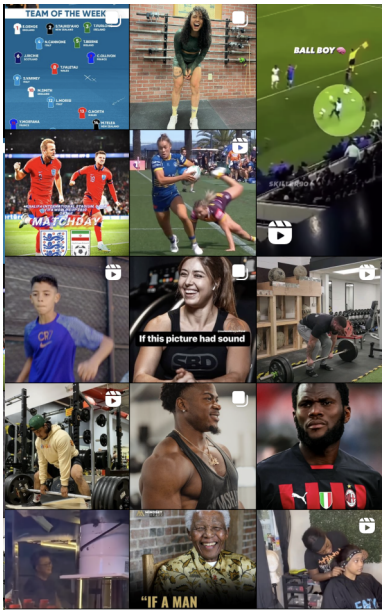

Figure 2: Another Instagram for you page

This for you page shows very different content to the previous one; there is a lot of sport/gym posts and absolutely no posts on food. This helps to show that bias exists in social media.
Proof: Let us assume that no bias exists in these examples. Then simultaneously, the following must be true:

- The posts in figure 1 is representative of the entire distribution of posts on the social media site.
- The posts in figure 2 is representative of the entire distribution of posts on the social media site.

This, trivially, can not be true due to the fact that the two pages show differing content. □

This project attempts to show how bias is affected by how a user interacts with social media. There are a couple of challenges to this project. The first being, even though we have a definition of bias, we need a way to measure it. This

involves creating a model to detect what topic a post is about and figuring out how to identify the distribution of topics on a social media site. The second challenge is to setup rules that represent different strategies users can use to interact with social media. We can then implement these rules and analyse the change in topics over time.

1.1 Related work

1.1.1 Pythia - Litou and Kalogeraki (2017)

Pythia is an automated system for short text classification. It makes use of Wikipedia structure and articles to identify topics of posts. Essentially, "Wikipedia contains articles organized in various taxonomies, called categories". Pythia then goes on to use this information as their training data as well as handling sparseness in posts on social media.

1.1.2 Topic tracking of student-generated posts - Peng et al. (2020)

This paper proposes a solution for determining valuable information/topics discussed in student forums on online courses. It uses a model called "Time Information-Emotion Behaviour Model" or otherwise called "TI-EBTM" to detect key topics discussions , keeping in mind the progress of time throughout the forum.
Although this paper specializes in academic online forums, the approaches made could be relevant and useful for this project.

1.1.3 Topic classification of blogs - Husby and Barbosa (2012)

This paper uses Distant Supervision - 'an extension of the paradigm used by (Snow et al. (2004)) for exploiting WordNet to extract hypernym (is-a) relations between entitities' - to get training data via Wikipedia articles. Then trains their own designed model on this data to be able to classify topics via a multi-class recognition model (69% accuracy) and via a binary classification model (90% accuracy).

1.2 Objectives

This project will require a topic detection model to be built and to perform some data analysis. An extension for this project will be to create a Chrome Extension.

Below are the objectives for this project (Objectives in green are completed):

- Build a topic detection model
  - Research possible models
  - Gather data for model training
  - Iterate over making and training models
  - Decide chosen model
  - Train model on larger dataset
  - Assess models accuracy and performance
  - Include other information in the model (e.g. image/media attached in posts)
- Data analysis
  - Determine the distribution of topics on social media - aka the "norm"
  - Create a set of rules to represent different strategies users can use to interact with social media
  - Implement these rules and analyse the change in topics over time
  - Compare the results of the rules to the "norm"
    - This may require the use of a mathematical model to determine differences over time-series data
- EXTENSION: Chrome Extension
  - Design UI for extension
  - Implement UI design
  - Gather data from social media sites from the extensions frontend

- Create an API that receives post information and returns a set of most represented topics
- Connect extension to API

2 Background

The below sections are areas of research that were completed to support the work in this project. Not all of the research areas are actively used in the project, but similar areas of work are also included in this section to add more context to the project.

2.1 Topic Modeling

Topic modeling is the process of extracting the topics of which a document represents. There are many different approaches to topic modeling, including: Latent Dirichlet Allocation (LDA), Formal Concept Analysis (FCA), and Latent Semantic Analysis (LSA). Below outlines a couple of the approaches that were considered for this project.

2.1.1 Latent Dirichlet Allocation (LDA) - Blei et al. (2003)

LDA is a generative probabilistic model for collections of discrete data. LDA makes the assumption that each document consists of a mixture of topics, and each topic consists of a mixture of words. Another assumption LDA makes is that the order of words in a document does not matter.
LDA is a good method for generating clusters of documents that are similar to each other. Then the clusters can be manually categorised into topics.

2.1.2 Formal Concept Analysis (FCA) - Priss (2006)

FCA is a mathematical framework for representing and reasoning about objects and their properties. The idea is to create a lattice (Partial Order) of concepts,

where each concept contains "Formal Objects" (Extension) and "Formal Attributes" (Intension).
The ordering of the lattice is determined by the following relation:

$$A \leq B \iff Extension(A) \subseteq Extension(B) \wedge Intension(B) \subseteq Intension(A)$$

Where A and B are concepts.
Let's show an example of a concept lattice. First lets define a set of objects and attributes:

	Sport	**Ball**	**Racket**	**Athletics**	**Game**
Tennis	X	X	X		X
Rugby	X	X			X
100m	X			X	
Shotput	X	X		X	
Chess					X

Table 1: Table of Objects and Attributes

We can then figure out all concepts (including implied concepts) by figuring out the extent and intent of each concept.

Using table 2 we can then create a concept lattice. the ordering of the lattice is determined by relation stated above.
When creating a concept lattice we also remove explicit relations if they are implied by other relations.

With the concept lattice in figure 3 we can now identify more generalized/specific topics; The more general topics are at the top of the lattice, and the more specific topics are at the bottom of the lattice.

This approach is looked at in this report as its a way of identifying topics and there relations to each other. However, the concept lattices can become very large and complex, and it is difficult to identify the most important topics. Another paper, that uses a similar approach to this (CigarrÃ¡n et al. (2016)) filters the attributes used to create a smaller, more manageable concept lattice.

Concept	**Extent**	**Intent**
Sport	Tennis, Rugby, 100m, Shotput	Sport
Ball	Tennis, Rugby, Shotput	Sport Ball
Racket	Tennis	Sport, Ball, Racket, Game
Athletics	100m, Shotput	Sport, Athletics
Game	Tennis, Rugby, Chess	Game
Sport ∩ Game (C1)	Tennis, Rugby	Sport, Ball, Game
Ball ∩ Athletics (C2)	Shotput	Sport, Ball, Athletics
Racket ∩ Athletics (C3)	∅	Sport, Ball, Racket, Athletics, Game

Table 2: Table of Extension/Intension

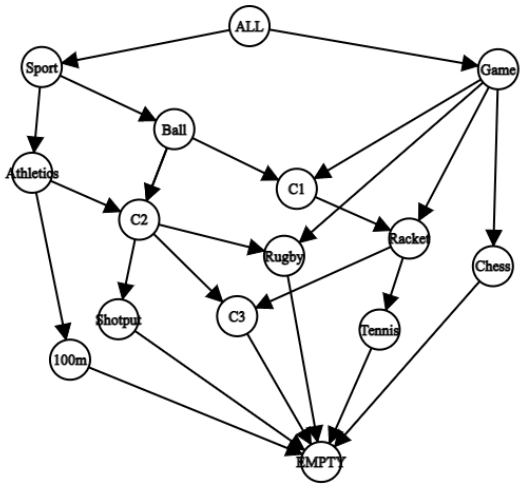

Figure 3: Concept Lattice

2.1.3 BERT and RoBERTa

BERT is a language model that was developed by Google. It is a bidirectional transformer model that uses a masked language model and a next sentence prediction task to train the model Devlin et al. (2018). The model is is trained in two stages. First, the model is pre-trained on a large corpus of text, and then the model is fine-tuned on a specific task.
The pre-training phase involves masking out random words in the corpus and training the model to predict the masked words, as well as next sentence prediction. The paper published by google: **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding** Devlin et al. (2018), goes into detail on the BERT architecture and training methodologies.
The fine-tuning phase involves training the model on a specific task. A further output layer can be added to the model, and then basic multi-class classification can be performed using a set of labelled data.

The paper written by Glazkova (2021) analyses BERT (as well as modified BERT models such as RoBERTa) and how they can be used for text classification. The data used in this paper is a set of scientific papers that are classified into 7 different categories.

The paper shows that using a Feedforward Neural Network (FNN) on top of BERT can achieve a 91.76% accuracy on the dataset. This is a very high accuracy, and shows that BERT can be used for text classification.

2.2 Data Analysis

2.2.1 Dynamic Time Warping - Müller (2007)

When analysing bias, it is required that we analyse the data compared to a baseline. On top of this, our analysis wants to see the change in bias over time. This is where Dynamic Time Warping (DTW) comes in. DTW is a technique

that allows us to compare two time series, even if they are not synchronised. This is useful for our analysis, as it will allow us to compare the data time-series to the baseline time-series, even if they are not synchronised.

The idea of DTW is to establish a function $c : A, B \rightarrow \mathbb{R}$, where A and B are two points in time series'. c is a function that assigns a cost noting the difference between the two points.
We can apply this function between all points in both time series. This will give us a matrix of costs.
The goal of DTW is to find the path through the matrix that has the total lowest cost. note, the path must make progress to the end at each step in the path. A valid path is shown in figure 4 and an invalid path is shown in figure 5.

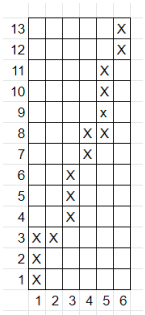

Figure 4: Valid DTW Path

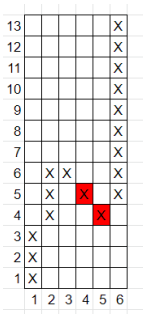

Figure 5: Invalid DTW Path

2.2.2 Search bias quantification - Kulshrestha et al. (2019)

This paper notes the difference between the bias in search results and the bias in a search query. The bias in a search query is the bias that is introduced by the user. Take for example, a user inputting a query for "World Cup Winners". This will ultimately lead to a lot of posts about football. But, this bias wasn't introduced by the search engine, it was introduced by the user.
Although, the specifics of input bias isn't useful for this project (we are not interested in search results) it did raise the question: Can we assume the distribution of posts on a social media platform is unbiased?

3 Progress

3.1 Weeks 1-2

During the first 2 weeks of term time was spent on the Specification as well as attempting to establish a strong base to the project in which we can build upon.

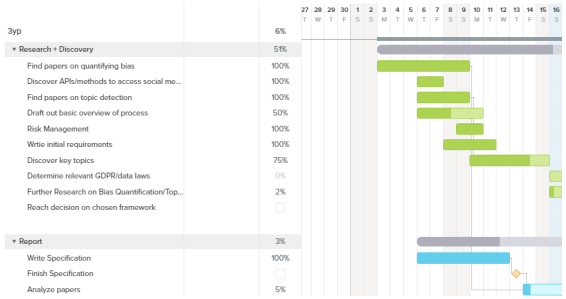

Figure 6: Timetable for weeks 1-2

As seen from this figure, most of the deadlines set for the first 2 weeks of the project. Some deadlines were missed Including: Drafting an overview of the process; Discovering key topics; and starting looking into GDPR and Data Protection laws. Although it would have been beneficial to have achieved the latter 2 in the list above, the first objective was probably infeasible for the first 2 weeks of the project - as further research is required to determine where the project was going to go.

A few extra tasks were completed early this week. =An API connection to Twitter was established as well as reading up on API access to Facebook and Instagram. As mentioned in my specification, Twitter will be the main use-case for this project but may look into using Facebook and Instagram as well.

Finally, further research into the papers laid out in the section 2 started. Specifically looking into Pythia and how they overcome the challenge of twitter

posts usually being short and not containing much information.

3.2 Weeks 3-4

During weeks 3 and 4, time was spent time was spent further working on establishing the goals of the project. It was decided that the project would be split into 2 parts. The first part would be to create a system that can be used to analyse twitter posts and determine the topic of the post. At this point, the chosen method was to go with a similar method used in Pythia Litou and Kalogeraki (2017) (Although this changes in weeks 5-6). The latter stages of the project would involve using the system and analyse how different social media "strategies" affect the topical bias of a users feed.

Below is the timetable for weeks 3-4.

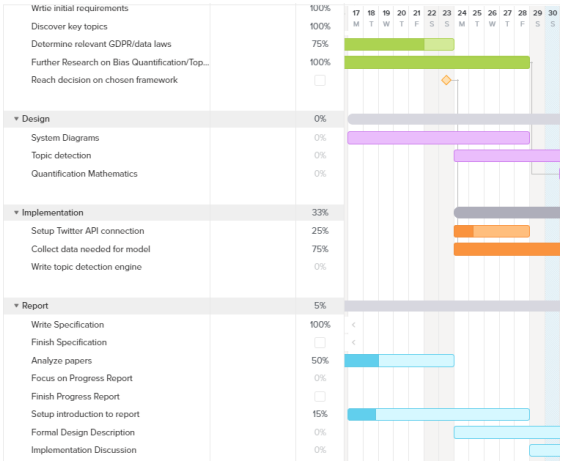

Figure 7: Timetable for weeks 3-4

As seen from the timetable above, deadlines set for weeks 3-4 were not completely met; due to needing to complete more research on other methods of topic detection (such as Formal Concept Analysis, Clustering, and Latent Dirichlet Allocation), The week was spent deciding which of these methods would be preferred (As mentioned above).

On top of this, data was gathered from Wikipedia as it was required for the training of the system. This involved:

- Connecting to the Wikipedia API
- Searching for the decided upon categories
- Selecting 100 pages for each category
- Performing some preprocessing on the data
    1. removing punctuation
    2. removing numbers
    3. removing excessive whitespace (leaving only spaces)
    4. removing Stopwords

def: Stopwords are words that are commonly used in the English language, but are uninformative Sarica and Luo (2021).

This process gave us a total of $24,000$ sentences at around $1,000$ per category.

3.3 Weeks 5-6

During weeks 5 and 6, time was focussed around BERT. BERT, which stands for Bidirectional Encoder Representations from Transformers, is a modern model which focusses on the task of language modelling. It is a pre-trained model which can be used to perform more precise tasks after being fine-tuned Devlin et al. (2019). Fine-tuning of the model is done by adding a classification layer on top of the model and training the model on a specific task.

The model was fine-tuned on the Wikipedia data gathered in the previous weeks. The model was trained on 24 categories, including: Politics, Sports, Food, etc. The model was trained for 6 epochs, with a batch size of 32.

The single output layer used the softmax activation function. The model was trained using the Adam optimizer with the sparse categorical cross entropy loss function.

The data was split into 80% training data and 20% testing data. The model was trained on the training data and then tested on the test data. Initially, the model was not very impressive, only achieving around 40% accuracy.

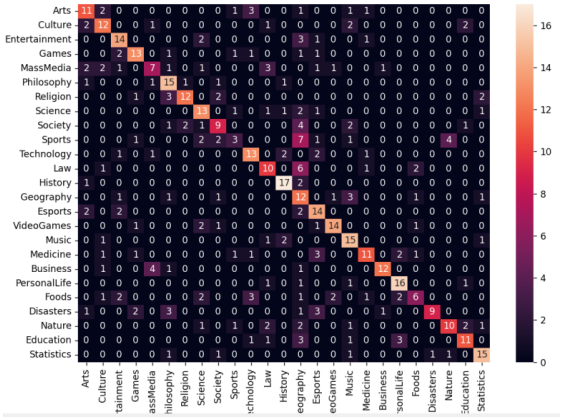

Figure 8: Confusion matrix on the Wikipedia data

Further analysis of the data showed that the data was of a poor quality; even after the processing, the data still contained meaningless sentences. Take for example: "51 (1): 209-220.", there is no way of identifying the topic of this sentence. On top of this, the data seemed to be too formal to be used for analysing social media. Because of these facts, the decision was made to not use Wikipedia data for the training and instead use Reddit data.
Reddit data was chosen as it is a more informal platform and is more likely to contain sentences that are formatted similarly to other social media sites.
The same process as laid out in the previous weeks was followed. The data was gathered, preprocessed. The model was trained and tested and with the reddit data, had a much higher testing accuracy of around 60%.

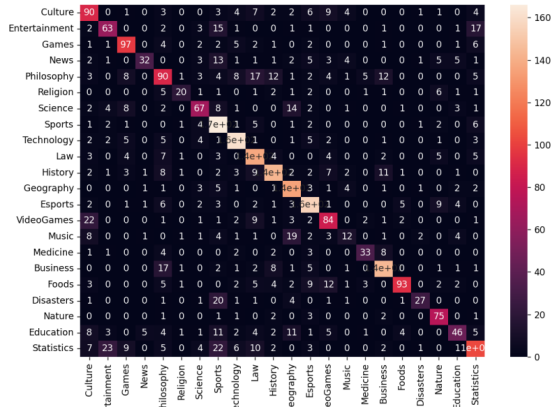

Figure 9: Confusion matrix on reddit data

3.4 Weeks 7-8

Little progress was made during weeks 7 and 8. This was primarily due to the fact the end of term was approaching and there were a lot of deadlines to meet. Mostly, these weeks was spent writing this progress report, and doing some more research into LDA and FCA. rotating

4 Project management

4.1 Initiation

The goal for the initiation of this project was to show the need for the project as well as setup a strong plan for the project. This is best achieved by creating a Project Initiation Document (PID).
A PID should contain the following information:

- Project overview
- Project objectives
- Project scope

- Project deliverables
- Project constraints
- Project assumptions
- Project risks
- Project budget
- Project schedule

The specification document was created to fulfil the requirements of the PID, and can be found in the appendix.

4.2 Planning

The planning for this project was also completed in the specification document. The method used for creating the plan was not optimal. This is due to the fact it was made completely on guess work. Although it is common practice for developers to take into account their previous experiences when creating a plan, it is not the best approach for this project due to the lack of experience available. This was ultimately the reason for the plan being changed during the project; the plan was too fast paced and did not allow for enough time to complete the tasks.

The plan could have been further improved by creating a Work Breakdown Structure (WBS). This would have given a broken down view of the project and have allowed for better estimation of the time required for work package. For the updated plan, a WBS was created.

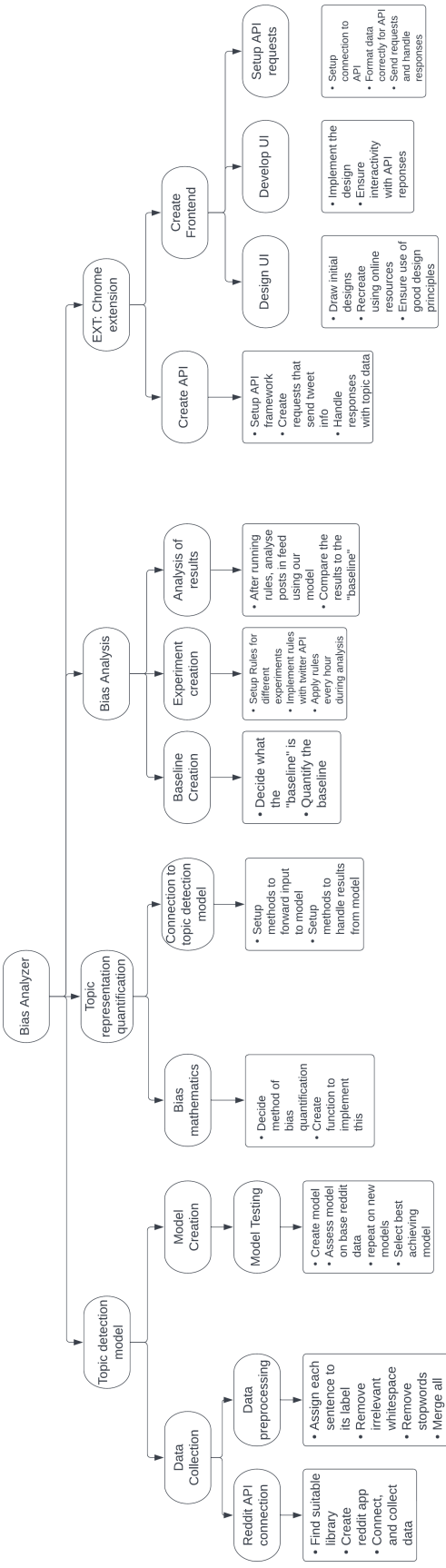

Figure 10: Work Breakdown Structure

4.3 Execution

An agile approach has been chosen for the execution of this project. This is due to the fact that the project involves a lot of experimentation. The initial stages of implementation required setting up different neural networks and comparing their performance. This is best achieved by using an agile approach as agile allows for changes during development **?**

The chosen agile approach is an individual version of Scrum. A set of goals is established every 2 weeks and then the progress is evaluated at the end of the 2 weeks. This is most evidently shown by section 3 of this report.

Implementation was only started recently. The plan is to stick to an agile approach for the model development. The data analysis part of the project will be completed using a waterfall approach. Firstly, a plan will be developed and then the analysis will be performed.

4.4 Timetable

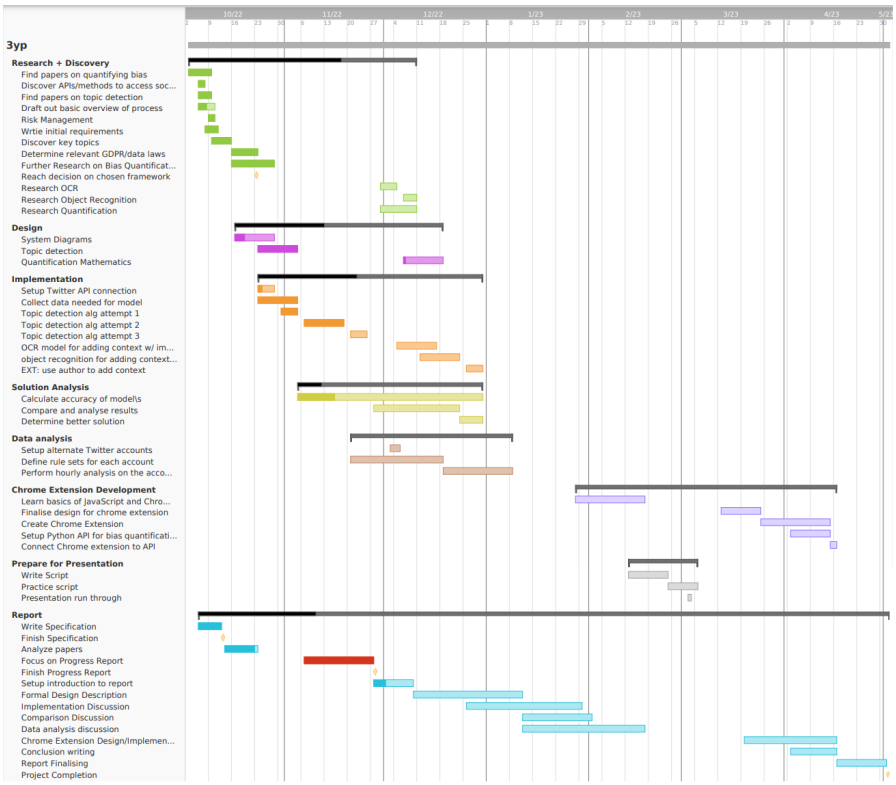

Figure 11: Updated Timetable for Rest of Project

References

Blei, David M & Ng, Andrew Y & Jordan, Michael I. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

CigarrÃ¡n, Juan & Ãngel Castellanos, & GarcÃa-Serrano, Ana. A step forward for topic detection in twitter: An fca-based approach. *Expert Systems with Applications*, 57:21–36, 2016. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2016.03.011. URL `https://www.sciencedirect.com/science/article/pii/S0957417416301038`.

Devlin, Jacob & Chang, Ming-Wei & Lee, Kenton & Toutanova, Kristina N. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018. URL https://arxiv.org/abs/1810.04805.

Devlin, Jacob & Chang, Ming-Wei & Lee, Kenton & Toutanova, Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. URL http://arxiv.org/abs/1810.04805. arXiv:1810.04805 [cs].

Glazkova, Anna. Identifying topics of scientific articles with bert-based approaches and topic modeling. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 98–105. Springer, 2021.

Husby, Stephanie & Barbosa, Denilson. Topic classification of blog posts using distant supervision. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 28–36, 2012.

Instagram, . How Instagram determines which posts appear as suggested posts | Instagram Help Centre. URL https://help.instagram.com/381638392275939.

Kulshrestha, Juhi & Eslami, Motahhare & Messias, Johnnatan & Zafar, Muhammad Bilal & Ghosh, Saptarshi & Gummadi, Krishna P. & Karahalios, Karrie. Search bias quantification: investigating political bias in social media and web search. *Information Retrieval Journal*, 22(1):188–227, Apr 2019. ISSN 1573-7659. doi: 10.1007/s10791-018-9341-2. URL https://doi.org/10.1007/s10791-018-9341-2.

Litou, Iouliana & Kalogeraki, Vana. Pythia: A system for online topic discovery of social media posts. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, pages 2497–2500, 2017. doi: 10.1109/ICDCS.2017.289.

Müller, Meinard. *Information Retrieval for Music and Motion*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. ISBN 978-3-540-74047-6 978-3-540-74048-3. doi: 10.1007/978-3-540-74048-3. URL http://link.springer.com/10.1007/978-3-540-74048-3.

Peng, Xian & Han, Chengyang & Ouyang, Fan & Liu, Zhi. Topic tracking model for analyzing student-generated posts in spoc discussion forums. *International Journal of Educational Technology in Higher Education*, 17 (1):35, Sep 2020. ISSN 2365-9440. doi: 10.1186/s41239-020-00211-4. URL `https://doi.org/10.1186/s41239-020-00211-4`.

Priss, Uta. Formal concept analysis in information science. *Annu. Rev. Inf. Sci. Technol.*, 40(1):521–543, 2006.

Sarica, Serhad & Luo, Jianxi. Stopwords in technical language processing. *Plos one*, 16(8):e0254937, 2021.

Shin, Donghee. How do users interact with algorithm recommender systems? the interaction of users, algorithms, and performance. *Computers in Human Behavior*, 109:106344, 2020. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2020.106344. URL `https://www.sciencedirect.com/science/article/pii/S0747563220300984`.

Snow, Rion & Jurafsky, Daniel & Ng, Andrew. Learning syntactic patterns for automatic hypernym discovery. *Advances in neural information processing systems*, 17, 2004.