A close-up of a red sign

Description automatically generated

Cardiff University

School of Computer Science and Informatics

CM3203 – Individual Project Report

Understanding the Effects of Fine-Tuning on NLP Models for Bias Detection in News Articles

Student: Jose Oscar Dos Santos Nunes

Number: 21055807

Supervisor: Dr [Fernando Alva Manchego](https://pats.cs.cf.ac.uk/!user_info?u=scmfa1)

Moderator: Dr [Neetesh Saxena](https://pats.cs.cf.ac.uk/!user_info?u=scmns2)

Abstract

The shift to digital media has transformed the consumption of news, with over 90% of adults in developed nations accessing news online. This change has led to a large increase in exposure to biased content, magnified by algorithms prioritizing user engagement over the accuracy of information. Addressing this issue is crucial for maintaining an informed public. Bias within digital news is a significant issue, distorting public perception and featuring societal divisions. The spread of biased content through digital platforms has worsened this problem, with 68% of Americans viewing biased reporting as a major concern. This project examines the effects of different fine-tuning strategies on pre-trained NLP models to improve bias detection in news articles. By fine-tuning these models on curated datasets and evaluating their performance, the study aims to identify the most effective approaches. The project also explores model performance on large text datasets to gauge their practical applicability. The methodology involved training and testing models using Hugging Face datasets, such as BABE-v3 and pranjali97, with an 80/20 train-test split. The project included a preliminary experiment on cross-training models, followed by a refined experiment focusing on the effects of fine-tuning. Evaluation metrics included precision, recall, F1 score, and accuracy. The D4DATA-on-BABE model showed the highest performance, with precision at 79.3%, recall at 84.5%, F1 score at 81.8%, and accuracy at 75.6%. Analysis across various categories provided insights into the models' strengths and weaknesses, demonstrating the importance of well-curated datasets. Future efforts will focus on expanding datasets, exploring multilingual capabilities, and developing custom models. Potential applications include real-time bias detection tools, such as a Chrome extension and integration into word processors for real-time feedback. This work aims to continually promote unbiased media consumption and improve public discourse.

Table of Contents

[Useful links 5](#_Toc166867619)

[Original models 5](#_Toc166867620)

[Short text datasets 5](#_Toc166867621)

[Long Text Datasets 6](#_Toc166867622)

[GitLab 6](#_Toc166867623)

[Hugging Face Model Repos 6](#_Toc166867624)

[Chapter 1: Introduction 7](#_Toc166867625)

[1.1. Motivation 7](#_Toc166867626)

[1.2. Problem 7](#_Toc166867627)

[1.3. Objectives 8](#_Toc166867628)

[1.4. Contributions 8](#_Toc166867629)

[Chapter 2: Background 9](#_Toc166867630)

[2.1. What is bias? 9](#_Toc166867631)

[2.2. How is bias demonstrated in text 9](#_Toc166867632)

[2.3. Effects of bias 10](#_Toc166867633)

[2.4. Bias in news reporting 11](#_Toc166867634)

[Chapter 3: Current technologies and related works 12](#_Toc166867635)

[3.1. Hugging Face Hub 12](#_Toc166867636)

[3.2. Large language models 12](#_Toc166867637)

[3.2.1. LLM chatbots (ChatGPT/GEMINI) 13](#_Toc166867638)

[3.3. News Bias Group 13](#_Toc166867639)

[Chapter 4: Methodology 15](#_Toc166867640)

[4.1. Dataset and data handling 15](#_Toc166867641)

[4.2. Standard practices/plan for evaluations 15](#_Toc166867642)

[Chapter 5: Preliminary experiment 17](#_Toc166867643)

[5.1. Introduction 17](#_Toc166867644)

[5.2. Models 17](#_Toc166867645)

[5.3. Datasets 17](#_Toc166867646)

[5.4. Training plan 17](#_Toc166867647)

[5.4.1 Cross-Training Strategy 17](#_Toc166867648)

[5.4.2. Training Procedures 18](#_Toc166867649)

[5.4.3. Training Configurations 18](#_Toc166867650)

[Chapter 6: Outcomes of Preliminary Experiment 19](#_Toc166867651)

[6.1. inadequate planning of experiment 19](#_Toc166867652)

[6.2. Dataset Structure and Model Training 19](#_Toc166867653)

[6.3. Lessons Learned 19](#_Toc166867654)

[6.4. Conclusion 20](#_Toc166867655)

[Chapter 7: Assessment Plan for primary experiment 21](#_Toc166867656)

[7.1. Introduction 21](#_Toc166867657)

[7.2. Models 21](#_Toc166867658)

[7.3. Datasets 21](#_Toc166867659)

[7.4. Model Training Approach 22](#_Toc166867660)

[7.5. Evaluation Metrics 23](#_Toc166867661)

[7.6. large data evaluations 23](#_Toc166867662)

[7.7. Tools and Software 24](#_Toc166867663)

[7.8. Ethical considerations 24](#_Toc166867664)

[7.9. Final Remarks 24](#_Toc166867665)

[Chapter 8: Implementation of primary experiment 26](#_Toc166867666)

[8.1. Overview 26](#_Toc166867667)

[8.2. Experiment Setup 26](#_Toc166867668)

[8.2.1. Hardware Specifications 26](#_Toc166867669)

[8.2.2. Operating System 26](#_Toc166867670)

[8.2.3. Software and Libraries 26](#_Toc166867671)

[8.2.4. Development Environment 26](#_Toc166867672)

[8.2.5. Version Control 26](#_Toc166867673)

[8.3. Model Training and Evaluations 27](#_Toc166867674)

[8.3.1. Data Handling and Pre-processing 27](#_Toc166867675)

[8.3.2. Model Configuration and Training 27](#_Toc166867676)

[8.3.3. Training arguments 27](#_Toc166867677)

[8.4. Evaluations and Metrics 28](#_Toc166867678)

[8.5. Challenges and Solutions 28](#_Toc166867679)

[8.6. Conclusion 28](#_Toc166867680)

[Chapter 9: Results 29](#_Toc166867681)

[9.1. Overall model performance 29](#_Toc166867682)

[9.1.1. Summary Statistics 29](#_Toc166867683)

[9.1.2. Models Performance Overview 30](#_Toc166867684)

[9.2. Topic and Type model performance(grouped analysis). 30](#_Toc166867685)

[9.2.1. Important preliminary information 30](#_Toc166867686)

[9.2.2. Model Performance Across Various Categories and Topics 31](#_Toc166867687)

[9.2.3. Key Observations 32](#_Toc166867688)

[Chapter 10: Analysis of results 33](#_Toc166867689)

[10.1. General model results 33](#_Toc166867690)

[10.1.1. General performance statistics 33](#_Toc166867691)

[10.1.2. Further evaluations of general performance. 35](#_Toc166867692)

[10.2. Models results sorted by type and topic. 35](#_Toc166867693)

[10.2.1. Bar graphs 35](#_Toc166867694)

[10.2.2. Confusion matrices 38](#_Toc166867695)

[10.2.3. Impact of Training Sets on Performance 40](#_Toc166867696)

[10.3. Analysis Summary 41](#_Toc166867697)

[Chapter 11: Plan of full news articles evaluation 42](#_Toc166867698)

[11.1. Objective 42](#_Toc166867699)

[11.2. Data Preparation 42](#_Toc166867700)

[11.3. Training plan 42](#_Toc166867701)

[11.4. Majority Vote labelling 42](#_Toc166867702)

[11.5. Model Prediction 43](#_Toc166867703)

[11.6. Performance metrics 43](#_Toc166867704)

[11.7. Expected Outcomes 43](#_Toc166867705)

[Chapter 12: Results and evaluations of full articles 44](#_Toc166867706)

[12.1. Results 44](#_Toc166867707)

[12.2. Evaluation 45](#_Toc166867708)

[12.2.1. Performance deviation graphs 46](#_Toc166867709)

[12.3. Final thoughts 49](#_Toc166867710)

[12.3.1. Model Performance Summary 49](#_Toc166867711)

[12.3.2. Areas for Improvement 49](#_Toc166867712)

[12.3.3. Concluding Remarks 49](#_Toc166867713)

[Chapter 13: Conclusion 50](#_Toc166867714)

[13.1. Problem and Aim Recap 50](#_Toc166867715)

[13.2. Concluding on objectives 50](#_Toc166867716)

[13.3. Concluding thoughts 51](#_Toc166867717)

[Chapter 14: Future work 52](#_Toc166867718)

[14.1. Furthering this experiment 52](#_Toc166867719)

[14.1.1. More accessible work 52](#_Toc166867720)

[14.1.2. Dataset improvements 52](#_Toc166867721)

[14.1.3. Develop custom model. 53](#_Toc166867722)

[14.1.4. LLM experimentation 53](#_Toc166867723)

[14.2. Potential application cases 53](#_Toc166867724)

[14.2.1. Chrome extension for real time detection 53](#_Toc166867725)

[14.2.2. Bias detection in word processors 54](#_Toc166867726)

[Chapter 15: Reflection on learning 55](#_Toc166867727)

[15.1. Challenges and Solutions 55](#_Toc166867728)

[15.2. Learning Outcomes 55](#_Toc166867729)

[15.3. Methodology and Implementation 55](#_Toc166867730)

[15.4. Results and Future Work 55](#_Toc166867731)

[15.5. Personal Reflection 56](#_Toc166867732)

[15.6. Conclusion 56](#_Toc166867733)

[References 57](#_Toc166867734)

**Figures**

[Figure 1: Original models on babe-v3. 22](#_Toc166867735)

[Figure 2: Original models on pranjali97. 22](#_Toc166867736)

[Figure 3: Models trained on Babe trained on pranjali97. 23](#_Toc166867737)

[Figure 4: graph of overall model performance across all metrics. 33](#_Toc166867738)

[Figure 5: Recall by Category (D4DATA-on-BABE) 35](#_Toc166867739)

[Figure 6: Recall by Category (D4DATA-on-BABE-on-PRANJALI) 36](#_Toc166867740)

[Figure 7: Recall by Category (VALURANK-on-BABE-on-PRANJALI) 37](#_Toc166867741)

[Figure 8: Confusion Matrix for D4DATA-on-BABE. 38](#_Toc166867742)

[Figure 9: Confusion matrix for D4DATA-on-BABE-on-PRANJALI. 39](#_Toc166867743)

[Figure 10: Confusion matrix for D4DATA-on-BABE-on-PRANJALI. 40](#_Toc166867744)

[Figure 11: Model performance on full News Articles 45](#_Toc166867745)

[Figure 12: Accuracy Deviation Graph 46](#_Toc166867746)

[Figure 13: F1 Deviation Graph 47](#_Toc166867747)

[Figure 14: Recall Deviation Graph 48](#_Toc166867748)

[Figure 15: Precision Deviation Graph 48](#_Toc166867749)

**Tables**

[Table 1: overall model results (rounded to 3 significant figures) 29](#_Toc166867750)

[Table 2: Counts of Types and Topics on Babe-v3 test and train split. 30](#_Toc166867751)

[Table 3: model results for type and topic (rounded to 3 significant figures) 31](#_Toc166867752)

[Table 4: Model Results of full article assessments 44](#_Toc166867753)

# Useful links

Git lab repo: <https://git.cardiff.ac.uk/c21055807/dissertation>

## Original models

* D1V1DE/bias-detection: <https://huggingface.co/D1V1DE/bias-detection>
* d4data/bias-detection-model: <https://huggingface.co/d4data/bias-detection-model>
* valurank/distilroberta-bias: <https://huggingface.co/valurank/distilroberta-bias>

## Short text datasets

* mediabiasgroup/BABE: <https://huggingface.co/datasets/mediabiasgroup/BABE>
* pranjali97/Bias-detection-combined: <https://huggingface.co/datasets/pranjali97/Bias-detection-combined>

## Long Text Datasets

* All The News: <https://www.kaggle.com/datasets/davidmckinley/all-the-news-dataset>

## GitLab

* Project GitLab: <https://git.cardiff.ac.uk/c21055807/individual_project>

## Hugging Face Model Repos

All models used and finetuned can be found uploaded here.

* Hugging face repo: <https://huggingface.co/Onunes>

# Chapter 1: Introduction

## 1.1. Motivation

In recent years there has been an obvious shift in the ways that news is consumed, most likely driven by the rise of the internet and the growing popularity of social media. Reuters Institute found that over 90% of adults in most developed nations access news online, with a majority of that under the age of 35 using platforms such as Facebook, Twitter, and TikTok (Reuters Institute, 2023). This transition from the traditional forms such as print, and broadcast media aligns with the more technological lifestyle of the digital generation with their offering of on-demand news content.

This change has major implications for society. Unlike conventional media, internet platforms can operate with low regulation and lack rigorous journalistic standards that are expected with traditional news (Reuters Institute, 2023). Official news and user-generated material are often mixed, making it difficult to distinguish between that which is verifiable and misleading information. Algorithms aimed at increasing user interaction often prioritise popular, provocative, or emotionally charged information over accuracy (Pew Research Center, 2023).According to the Reuters Institute, algorithmic prioritisation can strengthen existing preconceptions, limit the variety of available information, and create "echo chambers" that limit exposure to alternative ideas/viewpoints (Reuters Institute, 2023).

Young people are far more likely to absorb news through social media making this issue far more prevalent for them. According to Pew Research Center, 53% of regular news consumers on social media are under 30, and this demographic is also more likely to distrust news, perceiving it as biased or overly negative (Pew Research Center, 2023). Furthermore, a large portion of these young users-about 37%-actively avoid news due to its perceived negative impact on their mood and well-being (Reuters Institute, 2023).This growing dependence on digital platforms for news has implications for public discourse and democracy, shaping not only individual opinions but also collective consciousness on a global scale (Pew Research Center, 2023).

## 1.2. Problem

The pervasive effect of bias in digital news is a complex issue with technological, ethical, and cultural implications. Unlike conventional media, internet platforms have substantially lesser regulatory control, which increases the prevalence and effect of biased reporting. According to the Pew Research Centre, 68% of Americans feel biased news has a substantial impact on public opinion. This problem is worsened by the speed and reach of digital journalism, which may quickly promote biases and disinformation, leading to an inaccurate view of reality for many people.

Commercial incentives driving digital platforms often result in a conflict of interest where financial benefits from high user engagement are places above a journalist’s responsibility to provide balanced and accurate reporting. For instance, platforms such as YouTube and TikTok have become increasingly influential as news consumption mediums, prioritizing content based on engagement rather than accuracy, which raises concerns about potential bias​ (Nic Newman, 2023)​.

Additionally, the algorithms that organise and control content delivery on these platforms are not themselves neutral. Designed by humans and trained on datasets that could include biased human decisions, these algorithms can unintentionally worsen existing biases. This circumstance, known as algorithmic bias, creates a feedback loop where biased content is more likely to be produced, shared, and consumed. The Reuters Institute reported that almost half of surveyed individuals feel concern that algorithmic personalization may cause them to miss important information and challenging viewpoints (Fletcher, 2023). furthermore, research has shown that approval for algorithmic news selection is far higher among those with higher levels of trust in news, indicating the complicated relationship between trust, news consumption, and algorithmic influence (Pascal Jürgens, 2022).

The aim is not only to identify and eliminate biases in text, but also understand and prevent the processes by which these biases emerge and spread across digital news. This demands an in-depth knowledge of the relevant technology, such as algorithm design and data training procedures, as well as human behaviour and societal structures. Addressing such issues involves developing strategies in accordance with ethical AI practices, increasing algorithm openness, and educating a more critical and informed public. This complete strategy is critical for ensuring that digital platforms contribute to a more balanced and truthful coverage of news, promoting a better society (Caitlin, 2023).

## 1.3. Objectives

The primary aim of this project is to create a comprehensive evaluation of the effects of fine tuning on hugging face LM models for bias detection. The specific objectives to achieve are as follows:

1. Curate a set of comprehensive and labelled data set to be used for the evaluations of models.
2. Select a range of models to be used as the base models for further evaluations.
3. Plan, test, and analyse a comparative assessment on short form text data to understand the effects of cross training between selected models and their respective training datasets. Based on results a further evaluation of best performing models will discuss performance on full news articles.
4. Explain how the differing model architectures, training datasets affect performance.

Please not that these objectives have changed from those seen within the initial plan. Through further research and planning of the project we realised the original objectives were unrealistic given the timeframe and realistic scale of the project. As such the objectives were changed from the development of our own custom model to an evaluative structure of existing models.

## 1.4. Contributions

This project provides accessible and easy-to-understand information for researchers and students interested in bias detection in digital media. By using a combination of pre-existing datasets and models from Hugging Face, it offers a practical evaluation of fine-tuning effects on NLP models for bias detection.  
A key contribution is the in-depth analysis of fine-tuning language models on curated datasets. This research offers insights into how different fine-tuning techniques and dataset structures impact model performance. The findings help demystify the process of model fine-tuning, making it more approachable for researchers new to this area.  
This project also serves as a reference for researchers aiming to create or improve their own bias detection models. Through providing detailed performance metrics, we allow for other to build upon findings and enhance their understanding of model behaviours. Furthermore the results of this not only apply to bias detection but through understanding the evaluation researchers can apply the key conclusions into models for bias detection and other text classification models. The findings regarding fine tuning techniques and dataset content/structure are applicable to far more than bias.   
Overall, this project contributes to the field of NLP by providing more accessible insights, hopefully supporting further research and innovation in bias detection.

# Chapter 2: Background

## 2.1. What is bias?

Bias is an inherent aspect of human cognition and interaction. It refers to the predispositions or leanings, often subconscious, that shape our understanding, actions, and decisions. While commonly discussed in the context of news and media, bias extends far beyond influencing a wide range of areas including workplace environments, day to day life, and academic research. Recognising and understanding the various forms of bias is crucial for informed and rational decision making in all contexts (Psycology Today, NA).

Cognitive biases are ranging and can have varying impacts. For instance, confirmation bias describes the brain’s tendency to seek and focus on information that supports existing beliefs, while ignoring opposing evidence. This can lead to a skewed understanding of events or issues (Psycology Today, NA). Another example is the fundamental attribution error, where individuals are prone to attribute someone else’s actions to their personality traits, rather than considering situational factors of which may influence their behaviour (Psycology Today, NA).

These biases can often aid in quicker decision-making through simplifying complex information. However, this utility has a risk as this same bias can lead to errors in judgement and often sustain prejudices and unfair practices. This dual nature has been heavily researched within psychology, highlighting their role not only as aids in efficient decision making, but also as contributors to systematic mistakes and biases (Johan. E. (Hans) Korteling, 2023).

In conclusion, understanding bias is important in many areas since it influences human behaviour as well as societal standards. A thorough examination of the many types of biases and their effects is critical for improving justice, accuracy, and inclusion in decision-making and social interactions.

## 2.2. How is bias demonstrated in text

Bias in text, whether it is in news articles, academic writing or in everyday communication, can appear in a range of subtle forms. Understanding these manifestations is key to recognising and analysing biased content. Some examples of how bias may be demonstrated are as follows:

1. **Word choice and tone:** The words and tone used tend to influence the readers perception of an issue. For example, using emotionally loaded words, such as referring to a protest as a riot, will sway a reader’s perception (Mastrine, How to Spot 16 Types of Media Bias , 2019).
2. **Framing of information:** The framing of information involves manipulating the context around an argument to promote a specific interpretation (Frasier, Unknown). This includes what the author emphasizes or downplays and can be misleading even if technically accurate.
3. **Source selection:** An author may intentionally/unintentionally select sources which predominantly reinforce their own opinion. This can lead to the author not becoming aware of counterarguments or perspectives other than their own (Iowa State University Library Instruction, 2021).
4. **Perspective and focus:** The perspective from which a text is written may also introduce bias. For instance, writing about topics from a purely western perspective on a global issue may lead to a skewed representation of all facts and can overlook complexities and viewpoints of other cultures (Frasier, Unknown).
5. **Use of propaganda techniques:** Deliberate use of propaganda techniques such as intentionally using a combination of the pre explained examples, is a very clear demonstration of bias in text. These techniques can include inflammatory language, misleading contrasts, and framing bias​ (Frasier, Unknown).

By being aware of the methods in which bias can be demonstrated, readers can develop a more critical eye when consuming information. This awareness is crucial for discerning the reliability and fairness of the information they are presented with, and once achieved will allow for a more informed and balanced understanding of the world.

## 2.3. Effects of bias

Given its impact within society, public perception, and political processes, the impacts of media bias are an important concern in today's culture. Academic research by OUP (Oxford University Press) shows how biased news may impact public views by frequently instilling irrational fears that contradict objective world trends; news coverage often leads to a ‘mean world syndrome’. The mainstream media's tendency to exaggerate negative events can create a sense of fear and insecurity among their consumers (Oxford University Press, 2022).

In addition, to create a sense of fear and insecurity, media bias also plays a pivotal role in shaping political opinions and behaviours. News outlets with a specific political leaning often influence their audiences’ views, pushing the consumer towards a particular end of the political spectrum. The influence of this bias is especially prevalent within media echo chambers (An echo chamber is an environment where a person only encounters information or opinions that reflect and reinforce their own). (CGF GLOBAL, NA).

Furthermore, the perception of media bias significantly impacts public trust in the respective outlets. A report by Gallup suggests that most Americans acknowledge the critical role of media in democracy, however “More than eight in 10 Americans say the media bears "a great deal" (47%) or "a moderate amount" (36%) of blame for political division in this country.” (Gallup Inc, 2020).

In conclusion, media bias impacts society in important ways. It shapes how people view the world and influences their political beliefs. When the media portrays events with a certain slant, it can create a biased understanding of reality, which may lead to a decline in trust towards these media sources (Entman, 2007). As trust diminishes, societal divisions deepen, making constructive conversations more challenging and hindering effective public dialogue (Mark Jurkowitz, 2020).Ultimately, this can weaken the foundations of democracy, which relies on a well-informed public making decisions based on unbiased, accurate information. Thus, recognizing and addressing media bias is crucial for maintaining a healthy, democratic society (Stroud, 2011).

Top of Form

Bottom of Form

## 2.4. Bias in news reporting

Bias in news reporting significantly impacts how audiences perceive and understand the news. News organizations and journalists strive to adhere to a code of ethics to ensure fair and reliable reporting, but biases can still manifest in various ways (Entman, 2007). These biases may not only arise in the interpretation and presentation of facts, but also in the selection of stories and sources. For example, the decision about which stories to cover, the sources used, and how the information is displayed to the consumer, can all reflect a particular bias (Maxwell E. McCombs, 1972). The presence of such biases can lead to a skewed public perception, which emphasizes the need for critical media consumption skills among audiences (Rosenstiel, 2010).

An important aspect of understanding media bias is recognising the distinction between news gathering and news analysis. News gathering involves investigative word and fact-checking, while news analysis interprets the facts within a larger narrative. Newspapers such as The NY Times and the Wall Street journal engage in both activities. The bias often noticed in such outlets primarily pertains to the news analysis segment, influencing what they choose to cover and how they arrange the collected facts (Engle, 2024).

A good example for bias within news media can be seen with racial bias. Studies have shown that minorities are more often portrayed negatively in the news in comparison to others. This has been seen to influence public perception and thus societal attitudes. An example of this is to look at police stories where: African Americans have been overrepresented as perpetrators in crime news, whereas whites have been overrepresented as police staff (Wihbey, 2015).

The way media presents different stories may suggest various biases. For example, a heavy reliance on police enforcement narratives in reporting may unintentionally develop prejudices against specific communities. In foreign policy reporting, there may be an implicit assumption regarding the legitimacy of a country's aims, which may not always be supported by documented facts (FAIR, 2012).

Overall, bias in news is a complex issue that requires careful consideration and critical analysis by its consumers. By being aware of different forms of bias and understanding the distinction between accurate/factual reporting and analysis, readers can better navigate the range of media sources and gain a more balanced and informed perspective of different stories. Recognising the inherent biases in all forms of media, is crucial for a more comprehensive understanding of the news (Memmott, 2023) (Mastrine, What is Media Bias?, 2022).

# Chapter 3: Current technologies and related works

## 3.1. Hugging Face Hub

The Hugging Face organisation are leaders in AI research and have made significant contributions to NLP through their innovative transformer library. This library streamlines access to machine learning models, including popular ones such as BERT and GPT. These models are industry standards for NLP tasks including text categorization, sentiment analysis, and text generation, providing researchers with strong tools to solve challenging machine learning challenges.

Hugging Face also provides a dataset library to manage and access a wide array of datasets. This module simplifies data loading and pre-processing, which is crucial for successful model training and testing. The datasets have a wide variety of applications, from basic translation to more complex issues such as question answering and text summarising.

They have helped build a strong network of machine learning engineers. The platform encourages open collaboration and resource sharing, which has accelerated innovation and aided in the creation of new technologies. They offer forums for exchanging ideas, sharing results, and staying up to date on AI trends and breakthroughs.

Finally, Hugging Face is devoted to the ethical development of AI technology. It encourages the use of AI in a socially responsible and mutually beneficial manner. Hugging Face helps researchers uncover and minimise biases in AI models through their tools and guidelines. This focus on ethics has ensured that advancements in AI contribute positively to society and align with human moral standings.

To summarise, Hugging Face enhances AI research by providing necessary tools, encouraging collaboration, and emphasising ethical behaviours. Its contributions have not only accelerated innovation, but also ensured that AI breakthroughs are designed with social ideals in mind.

## 3.2. Large language models

Large language models (LLMs), such as GPT-4 and Gemini, represent a significant advancement in the field of AI and NLP. These models are built upon deep learning techniques and are trained on vast amounts of text data.

LLMs have been used for a wide range of tasks such as translating and summarising text, along with answering questions and writing code.

While they are not specifically designed for the purpose, due to the nature of the systems, they can be utilised for the detection and analysis of text. Bias be of different types e.g. gender, racial or cultural bias, and can be both explicit and implicit. As LLMs are trained on such vast quantities of data, it means that they can potentially be better equipped for identifying a wider range of biases a model for identifying one type of bias.

Differently to other NLP processes such as sentiment analysis, there is a far longer process for using LLMs in bias detection. They are often found as follows:

1. **Data Analysis:** LLMs can process extensive amounts of text to identify potentially biased language. This can include analysing word associations, sentence structure and the context of specific terms/phrases used (Aylin Caliskan, 2017).
2. **Pattern Recognition:** Through their training, LLMs learn to recognise patterns in language. This capability can be utilised to identify recuring themes that may suggest the presence of bias (Jacob Devlin, 2019).
3. **Contextual Understanding:** One key strength of LLMs is their ability to understand context. This is crucial in bias detection, as the context in which words are used can significantly alter their meaning and thus perception (Matthew E. Peters, 2018).
4. **Comparative Analysis:** LLMs can compare different text sources or segments within the same entity to identify variations that may indicate bias. For example, an LLM could be used to understand the description of similar achievements by different groups of people (Jieyu Zhao, 2018).

There is, however, one large risk with LLMs, that being that they themselves are subjectable to bias. All models run the risk of being biased through potentially being trained with biased/inaccurately labelled data. With LLMs, this isn’t typically a huge issue as they are often trained on data sources from ranging curators, whereas smaller niche models may be more impacted (Emily M. Bender, 2021).

### 3.2.1. LLM chatbots (ChatGPT/GEMINI)

LLM chatbots such as ChatGPT and GEMINI are not designed for media bias detection. However, their ranging capabilities in language understanding make them a useful assistance tool for bias detection.

These chatbots not only display an advancement in current technology but also as an insight to where AI, for tasks such as bias detection, may lead. These models offer a fast and in-depth analysis approach to understanding biases, however, it is crucial to acknowledge that while these AI tools are powerful, they are not specifically designed for bias detection and should thus be used in conjunction with other related tools to ensure the best results.

Unfortunately, LLM experimentation is outside the scope of this project as we are only working with language models. However, this topic is further discussed in the Future work section seen towards the end of this paper.

## 3.3. News Bias Group

The Media Bias Group is a network of researchers focused on working on understanding media bias from various perspectives including but not limited to computer science, economics, and politics. As an organisation, they aim to understand how media bias is perceived, represented, and influences decision-making. Their research spans many fields but includes developing methodologies for creating comprehensive data sets on media bias, and automated detection of biased language. Additionally, they have studies on the role of linguistic features in media bias and work on creating visualisations to represent media bias effectively.

The organisation is very present on the Hugging Face platform, offering a range of public datasets and models that can be used for both independent and collaborative research. The datasets are diverse and serve a range of purposes in the study and analysis of media content. For the case of this project, I have primarily been looking at “BABE” and “BABE -v3” (with “BABE -v3” being a larger version using data from other projects). I will elaborate further on the datasets in later sections of the project (see 4.1. Dataset and data handling).

While the datasets are useful for researchers, the most beneficial work done by the group comes with their models. The models developed by the Media Bias Group are primarily focused on analysing and detecting bias in media content. These models are based on advanced machine learning and NLP techniques. There is a focus on the [BERT](https://huggingface.co/docs/transformers/en/model_doc/bert) framework and its variants such as [RoBERTa](https://huggingface.co/docs/transformers/en/model_doc/roberta).

* **Bias Detection:** Many of the models are designed to detect bias in media. This includes identifying political biases, detecting loaded language, and recognising subtle forms of which may not be obvious. The models can analyse text to identify specific words, phrases and/or patterns that are indicative of potential bias.
* Text Classification: text classification is a common application within the models. They can categorize text. They care used to group text into various classifications based on the presence or type of bias. For example, their “RoBERTa” model is used to return a percentage of text which is deemed to be bias or neutral (as seen with the model [here](https://huggingface.co/mediabiasgroup/magpie-babe-ft?text=Since+then%2C+health+care+has+turned+out+to+be+a+very+strong+issue+for+Democrats%2C+who+campaigned+on+the+issue+aggressively+during+the+2018+midterms+and+enjoyed+a+net+gain+of+40+seats+in+the+U.S.+House+of+Representatives.)).

# Chapter 4: Methodology

## 4.1. Dataset and data handling

The technological components of the project have been developed using Jupyter notebooks due to their unique advantages over the traditional Python file formats. The cell structure of Jupyter allows for easy execution and modification of small code segments without executing the entire program, which is especially useful for data visualizations displayed directly within the notebook. This integrated approach likely speeds up code review and will enhance any readers understanding. Additionally, markdown cells enable developers to add comments and insights, improving communication and comprehension among team members. Overall, Jupyter notebooks provide an efficient environment for development, visualization, and collaboration.

For the project, I have decided to use CSV files for data storage. As the datasets are relatively small (around 4k rows as largest), CSV files are the most practical choice as they are lightweight and have minimal need for configuration and management in comparison to databases, many of the selected datasets are stored on hugging face as csv files so it requires minimal changes. While tools such as databases or excel files could be implemented, at this time their typical advantages are not as relevant in this case, so they have been decided against.

Additionally, CSV files integrate seamlessly with many of python’s data analysis libraries such as pandas. Acknowledging that calling a csv file every time data is needed, would be highly impractical I will be using data frames at runtime. Their implementation will allow for quick updates, manipulation, and analysis that is crucial for the project. Overall, the decision to utilise CSV files in this case lies in their balance of simplicity, functionality, and the requirements of the project.

One important note with the data storage is that a database would need to be implemented for potential further work on the project, such as using the models and tools discussed/researched to build and application or website for users to analyse their own text,

## 4.2. Standard practices/plan for evaluations

To keep the assessments fair, I will be using a standardised method for the evaluation of various models. Using the previously mentioned BABE-V3 dataset, I have created a train and test dataset. I am using the 80/20 split for train/test and this dataset will be created through a python script, randomly assigning rows into each of the datasets. Once this standardised test set is created, it will be used for all assessments where it is relevant. By keeping the testing set standardized, it ensures a fair evaluation of each of the models.

For each of the pretrained models they will be assessed in their initial state before any fine tuning, this will be a baseline for how well they can process data having been trained on a dataset especially made for its purpose. For the evaluation of the models, they will be tested using my testing section from the BABE-V3 dataset. The selected models will be fine-tuned on the BABE-v3 training split as well as the Pranjali97 dataset. This will result in 12 models that will be initially evaluated on the short form text data. These results will indicate 3 best performing models that are selected for further evaluations on the full news article data.

data to see how these models perform when tasked on a slightly alternative assessment.

# Chapter 5: Preliminary experiment

## 5.1. Introduction

This experiment aims to explore the effects of cross training between 3 models on 3 datasets. Each model will be trained on a combination of each dataset and then evaluated against a testing dataset. The results aim to show how dataset structure and model architecture affects performance and adaptivity.

## 5.2. Models

For this assessment, 3 models have been used. They are as follows:

* [D1V1DE/bias-detection](https://huggingface.co/D1V1DE/bias-detection) – This is a RoBERTa based model that has been trained on the mediabiasgroup/BABE.
* [d4data/bias-detection-model](https://huggingface.co/d4data/bias-detection-model) – A distilbert based model that has been trained on the [MBIC Dataset](https://github.com/Media-Bias-Group/Neural-Media-Bias-Detection-Using-Distant-Supervision-With-BABE)
* [valurank/distilroberta-bias](https://huggingface.co/valurank/distilroberta-bias) – Based on [distilroberta-base](https://huggingface.co/distilroberta-base), and trained on the [wikirev-bias](https://huggingface.co/datasets/valurank/wikirev-bias) dataset

## 5.3. Datasets

As previously states there are 3 datasets being used for this assessment. They are:

* [wikirev-bias](https://huggingface.co/datasets/valurank/wikirev-bias)
* [MBIC Dataset](https://github.com/Media-Bias-Group/Neural-Media-Bias-Detection-Using-Distant-Supervision-With-BABE)
* mediabiasgroup/BABE

These 3 datasets offer a range of different structures, size, and content. Given the differences in structures the results should offer a range of values allowing the accurate understanding of how dataset structure can affect accuracy of fine-tuned models.

## 5.4. Training plan

The training plan for the preliminary experiment is designed to assess the impact of cross-training three different models on three distinct datasets. This section outlines the methodology for training the models, the configurations used, and the rationale behind the chosen strategies.

### 5.4.1 Cross-Training Strategy

The training plan involves cross training each model on every combination of the three datasets. The objective is to understand how training on one dataset and fine-tuning on another affects the model's bias detection capabilities. The cross-training combinations are as follows:

1. D1V1DE/bias-detection:
   1. Trained on mediabiasgroup/BABE
   2. Fine-tuned on MBIC Dataset
   3. Fine-tuned on wikirev-bias
2. d4data/bias-detection-model:
   1. Trained on MBIC Dataset
   2. Fine-tuned on mediabiasgroup/BABE
   3. Fine-tuned on wikirev-bias
3. Valurank/distilroberta-bias:
   1. Trained on wikirev-bias
   2. Fine-tuned on mediabiasgroup/BABE
   3. Fine-tuned on MBIC Dataset

### 5.4.2. Training Procedures

Each model undergoes the following training procedures:

1. Initial Training:
   1. Models are first trained on their respective original datasets using standard training configurations.
   2. Training parameters such as learning rate, batch size, and the number of epochs is optimized to ensure stable and effective learning.
2. Fine-Tuning:
   1. After initial training, each model is fine-tuned on the other two datasets one after the other.
   2. Fine-tuning involves further training the model with a smaller learning rate to adjust the pre-trained model weights to the new dataset.
   3. This step is crucial for assessing how well the models adapt to new data and their ability to generalize across different datasets.

### 5.4.3. Training Configurations

The training configurations are standardized across all models to ensure fair comparisons. The configurations include:

* **Learning Rate**: A smaller learning rate (e.g., 5e-5) is used during fine-tuning to prevent drastic changes to the pre-trained model weights.
* **Batch Size**: A batch size of 16 is used to balance memory usage and training efficiency.
* **Number of Epochs**: The models are trained for 3 epochs during initial training and 2 additional epochs during fine-tuning.
* **Optimizer**: The AdamW optimizer is used for its effectiveness in handling sparse gradients.
* **Evaluation**: During training, models are evaluated on a validation set to monitor performance and prevent overfitting.

# Chapter 6: Outcomes of Preliminary Experiment

Unfortunately, this experiment was unsuccessful, and as such, there were no completed implementations, results, or analyses. Several factors contributed to this outcome, which will be explained in this section.

## 6.1. inadequate planning of experiment

A key factor in the unsuccess of this project was the lack of a clear plan for execution. Given that this was our first time experimenting with machine learning algorithms and NLP, we underestimated the challenges and the necessary skills required for this experiment. Several mistakes were made, including not checking the dataset structure, model architecture, and not fully planning out the evaluation process for these models. This lack of foresight led to significant setbacks and ultimately the failure of the experiment.

* **Underestimated Challenges**: The complexity of working with machine learning and NLP was not fully anticipated. The steep learning curve and the technical nuances were more demanding than initially expected.
* **Insufficient Skills**: The project required advanced knowledge and skills in data pre-processing, model training, and evaluation. The initial team composition lacked the necessary expertise in these areas.
* **Incomplete Planning:** Key steps in the project, such as detailed timelines, milestones, and contingency plans, were not adequately developed. This oversight led to a disorganized workflow and inefficient use of resources.

## 6.2. Dataset Structure and Model Training

As mentioned above, the datasets required for this assessment caused several issues. We assumed that all training datasets would be of a simple and similar format, which was an oversight.

* **Complex dataset structures**: while the babe dataset was easy to understand and simplistic in design. The Wiki-rev and MBIC datasets did had complex structures that we had not anticipated. These datasets contained varied libelling conventions, several different files of differing structures and contents.
* **Inconsistent Formats**: The datasets varied significantly in structure, labelling conventions, and size. This inconsistency made it difficult to standardize the pre-processing and training processes.
* **Quality Issues**: Some datasets contained noisy, complicated, or mislabelled data, which adversely affected the training process. The models struggled to learn from this data, resulting in poor performance.

## 6.3. Lessons Learned

Despite the failure, the experiment provided valuable insights and lessons that will be used to inform the following experiment.

* Comprehensive planning: The next experiment will benefit from a more thorough and thought-out plan. This will include a more comprehensive timeline, objectives, and a predefined evaluation plan.
* Dataset issues: As stated a key reason for the failure of the experiment came from the problems with regards to datasets. To solve this any potential datasets will be more rigorously checked and prepared before the start of experiment implementation.
* Skill Development: Through this experiment many of the necessary skills have been identified and learned. Due to this it should ensure that the future work will be completed at a faster rate and in more depth.

## 6.4. Conclusion

While the lack of success in this experiment has been unfortunate, we have gained much from the attempt as outlined in the “Lessons Learned” section. This work will not providing quantifiable results will be instrumental in the success of the following assessment.

# Chapter 7: Assessment Plan for primary experiment

## 7.1. Introduction

Due to the unsuccessful first experiment. we devised a new experiment to recreate all possible aspects and deal with the issues causing the lack of success. To keep similarity, we have kept the same 3 models and the Babe-v3 dataset. However, we have made the decision to drop any work with the MBIC and wikirev-bias datasets. In their place, a new one has been selected to be used as a second training dataset.

Within this assessment we are focusing on the small text evaluation coming from the “Babe-v3 test” split previously outlined.

## 7.2. Models

For this assessment, 3 models have been used. They are as follows:

* [D1V1DE/bias-detection](https://huggingface.co/D1V1DE/bias-detection) – This is a RoBERTa based model that has been trained on the mediabiasgroup/BABE.
* [d4data/bias-detection-model](https://huggingface.co/d4data/bias-detection-model) – A distilbert based model that has been trained on the [MBIC Dataset](https://github.com/Media-Bias-Group/Neural-Media-Bias-Detection-Using-Distant-Supervision-With-BABE)
* [valurank/distilroberta-bias](https://huggingface.co/valurank/distilroberta-bias) – Based on [distilroberta-base](https://huggingface.co/distilroberta-base), and trained on the [wikirev-bias](https://huggingface.co/datasets/valurank/wikirev-bias) dataset

## 7.3. Datasets

The data required for the project is split into two primary categories. This being unlabelled and labelled data.

The labelled data comes from 2 hugging face datasets, “mediabiasgroup/BABE-v3” and “pranjali97/Bias-detection-combined”. BABE-V3 is the primary dataset used for both training and the testing of a model’s performance, whereas pranjali97s data set is used for just fine tuning.

I will be using the [BABE-V3 dataset](https://huggingface.co/datasets/mediabiasgroup/BABE) found on Hugging Face for the vast portion of my project. This dataset has been curated by the previously mentioned news bias group. It is an improved dataset of the original BABE dataset and has been used to train several of their models. Whilst this dataset has many useful fields, such as listing of biased words present and links to news articles from where the text has been extracted, we are only using the text and label columns for the evaluations.

[Pranjali97s](https://huggingface.co/datasets/pranjali97/Bias-detection-combined) dataset has been selected for similar reasons. Whilst it is a dataset curated by a singular user, it has been selected due to its popularity within Hugging Face when compared to other bias dataset alternatives. It is a much simpler dataset containing only 3 fields; however, it holds the key fields of text and a label of bias being present or not.

The second group of data refers to the “[all the news](https://www.kaggle.com/datasets/davidmckinley/all-the-news-dataset/data)” data set found on Kaggle. This dataset has been selected to deal with the short comings of those mentioned previously and its fields of long text data. The Hugging Face datasets hold labelled data, where all text is short form (often a sentence or less). This dataset also holds full news articles, allowing for a more real-world application of the bias detection models. The drawback of this dataset is that it is unlabelled, however, this is dealt with through the implementation of a majority vote from the best performing models from the previous evaluations.

## 7.4. Model Training Approach

Instead of cross-validation, the model training approach is as follows:

* Single dataset fine tuning: All original models (outlined above) will be fine-tuned on my training set of the Babe-v3 dataset, as well as my training set of the pranjali97 dataset. (note: while D1V1DE was originally trained on Babe, Babe-v3 differs slightly so for the sake of consistently it has been included in the fine tuning)
* Secondary training: Once all models have been trained on both Babe-v3 and pranjali97, the models trained on Babe-v3 will be fine-tuned again on the pranjali97 dataset.

See below for a visualisation of training plan (the structure of the diagrams follows: data used for fine tunings; model being fine-tuned then the new created fine-tuned model) :

1. Figure 1 shows the process for creating the “D1V1DE-on-BABE”, “D4DATA-on-BABE”, “VALURANK-on-BABE” models.

A diagram of a software flow

Description automatically generated with medium confidence

Figure 1: Original models on babe-v3.

1. Figure 2 shows the process for creating the “D1V1DE-on-pranjali97”, “D4DATA-on-pranjali97”, “VALURANK-on-pranjali97” models.

A diagram of a software flow

Description automatically generated with medium confidence

Figure 2: Original models on pranjali97.

1. Figure 3 shows the secondary stage of training, where the models created in Figure 1 are further fine-tuned with the pranjali97 training set.

A diagram of a machine

Description automatically generated

Figure 3: Models trained on Babe trained on pranjali97.

## 7.5. Evaluation Metrics

This experiment will utilise standard evaluation metrics for machine learning models. These will include:

* **Accuracy:** This metric shows the proportion of correct predictions (both true positives and true negatives) relative to all evaluated cases. Although helpful for gauging overall performance, accuracy can sometimes be misleading, particularly in datasets with uneven class distributions.
* **Precision:** This evaluates the model's accuracy in identifying positive labels, crucial in bias detection. It calculates how frequently texts labelled as biased, are genuinely biased, using the ratio of true positives to all predicted positives (true positives and false positives). A model with high precision reduces the chances of mistakenly labelling unbiased content as biased.
* **F1-Score:** This metric provides a balanced view of both precision and recall, useful when seeking a balance between detecting as many positives as possible (high recall) and ensuring those positives are genuinely pertinent (high precision). The F1-score the harmonic mean of precision and recall, offers a more nuanced perspective on the model's performance in both areas.
* **Recall:** Measures the model's ability to identify all relevant positive cases. It indicates how well the model captures actual positive instances, ensuring no potential biases are overlooked.

These metrics collectively provide a robust framework for evaluating the models, aiding in their refinement for enhanced accuracy and practical utility in real-world settings where the stakes of incorrectly labelling or missing biased content are high. Their integrated use is crucial for developing models that effectively and reliably identify and categorize biased information.

## 7.6. large data evaluations

As previously mentioned, the long text data coming from the “all the news” data set is unlabelled. To overcome this issue such that the data can be used for evaluations, we employ a majority vote from the best performing models on the short data. These models were selected from the best performers when evaluated on the Babe-v3 test set.

This process/evaluation will be further explained in the Evaluation on full news articles section of this report.

## 7.7. Tools and Software

The evaluation will leverage:

* [**Scikit-learn**:](https://scikit-learn.org/stable/) For implementing machine learning metrics.
* [**TensorFlow**](https://www.tensorflow.org/) and [**PyTorch**](https://pytorch.org/): For modelling and evaluation of neural network-based classifiers.
* [Hugging Face **Transformers**](https://huggingface.co/docs/transformers/en/index): For handling NLP tasks and data processing efficiently.
* [**Matplotli**b](https://matplotlib.org/) and [**Seaborn**](https://seaborn.pydata.org/): For visualizing the results and fairness assessments.
* [**Jupyter notebooks**](https://jupyter.org/): Running within visual studio code, a Jupyter notebook is used for development of the assessment.

## 7.8. Ethical considerations

In our primary experiment which focuses on refining and applying bias detection models, traditional ethical considerations typical for human subjects are not required. Instead, the experiment adheres to ethical AI use principles using pre-trained models developed with ethical frameworks. Here are the key points:

1. **Use of Pre-Trained Models:** The models used in this experiment, like RoBERTa and DistilBERT, are refined versions of larger models created by well-regarded research groups. These models were built with a strong focus on ethical AI practices, including training data transparency, fairness, and inclusivity.
2. **Public Data Sources:** The data for this study comes from publicly available news articles. This approach ensures no confidentiality or privacy breaches, as the data does not contain sensitive or personally identifiable information.
3. **Focus on Bias Reduction:** The primary goal of using these models is to identify and lessen bias in text, aiming to reduce misinformation and enhance fairness in media reporting. This effort supports ethical AI use by striving to improve the neutrality and accuracy of public information.
4. **Transparency in Methodology:** The experiment upholds high transparency standards by clearly documenting the data sources, model configurations, and analytical methods used. This level of transparency promotes reproducibility and helps maintain trust in the research methods and findings.
5. **Adherence to Ethical AI Standards:** By using ethically developed models and refining them for specific uses, the project commits to maintaining ethical AI development and application standards. This includes regular assessments to prevent new biases during model refinement.

Thus, while specific ethical approvals are not necessary for this experiment, it is conducted with a proactive commitment to maintaining high ethical standards in AI usage, especially concerning bias detection and mitigation.

## 7.9. Final Remarks

This assessment is designed to test and evaluate the performance and adaptability of the models on varying data. It has been curated to deal with the issues present in the previous experiment and provides accurate and informative results and metrics. By employing quantitative metrics and qualitative analysis from the notebook, the plan aims to validate the models comprehensively and identify any necessary improvements.

# Chapter 8: Implementation of primary experiment

## 8.1. Overview

This section outlines the steps taken to prepare and execute primary experiment, focusing on the methodologies applied for data extraction and analysis. The experiment aimed to refine and validate the effectiveness of bias detection models on a curated set of data, addressing the shortcomings observed in previous experiments.

## 8.2. Experiment Setup

For the sake of reproducibility this section outlines all hardware and developmental settings used within the experiment. Please note that reproduced results may vary based on local hardware specifications and software versions.

### 8.2.1. Hardware Specifications

The experiments were conducted on a 2021 M1 MacBook Pro, equipped with an 8-core CPU, 8-core GPU, 16GB of unified memory, and 512GB SSD. This setup provided the necessary computational power and speed required for the data processing and model training tasks.

### 8.2.2. Operating System

The system was running macOS Big Sur version 11.2, optimized for M1 chips. This operating system version ensured smooth operation and compatibility with the development tools and libraries used in the experiments.

### 8.2.3. Software and Libraries

The experiment was implemented within a Jupyter notebook, running under a Python 3.10 virtual environment managed through Anaconda. This setup helped maintain project dependencies and facilitated reproducibility. For a detailed list of specific libraries and their versions, refer to the “requirements.txt” file within the project’s repository. This file includes all necessary Python packages, ensuring that other researchers can recreate the computing environment with ease.

### 8.2.4. Development Environment

All development work related to this experiment was carried out using Visual Studio Code, version 1.74.2. Visual Studio Code provided a robust, flexible platform for coding, with support for Python and integration with Jupyter notebooks, enhancing the efficiency of code development and testing.

### 8.2.5. Version Control

The project has been managed through Cardiff University’s GitLab (version unknown), which facilitated version tracking and collaboration. Using GitLab ensured that changes to the experiment’s code and documentation were systematically managed and documented, allowing for effective collaboration among multiple researchers.

## 8.3. Model Training and Evaluations

### 8.3.1. Data Handling and Pre-processing

We initiated the experiment by importing and pre-processing the datasets earmarked for model training and evaluations:

1. mediabiasgroup/BABE-v3: After importing, we cleansed the dataset by removing non-essential columns including news\_link, outlet, label\_opinion, and biased\_words. We generated binary labels based on the type of column, classifying texts into 'left', 'right', or 'centre'. The dataset was split into training (80%) and testing (20%) sets to ensure a robust training process.
2. pranjali97/Bias-detection-combined: This additional dataset was used primarily for fine-tuning purposes. It was loaded through the Hugging Face's datasets library, retaining only the critical fields necessary for bias detection tasks.

The prepared datasets were saved in CSV format to facilitate efficient loading during the training phases.

### 8.3.2. Model Configuration and Training

The training utilized three models, each pre-configured with state-of-the-art architectures available via the Hugging Face transformers library:

* D1V1DE/bias-detection (based on RoBERTa)
* d4data/bias-detection-model (based on DistilBERT)
* valurank/distilroberta-bias (based on DistilRoBERTa)

Each model underwent an initial fine-tuning phase on the BABE-v3 dataset. We employed the Trainer and TrainingArguments from the transformer’s library for this purpose, which provided comprehensive management of training operations including automatic data batching, sequence padding, and detailed logging.

Following this, a secondary fine-tuning stage was implemented using the pranjali97 dataset to evaluate the models’ adaptability and performance improvement when introduced to a varied textual dataset.

### 8.3.3. Training arguments

The training configurations are standardized across all models to ensure fair comparisons. The configurations include:

* **Learning Rate**: A smaller learning rate (e.g., 5e-5) is used during fine-tuning to prevent drastic changes to the pre-trained model weights.
* **Batch Size**: A batch size of 16 is used to balance memory usage and training efficiency.
* **Number of Epochs**: The models are trained for 3 epochs during initial training and 2 additional epochs during fine-tuning.
* **Optimizer**: The AdamW optimizer is used for its effectiveness in handling sparse gradients.
* **Evaluation**: During training, models are evaluated on a validation set to monitor performance and prevent overfitting.

## 8.4. Evaluations and Metrics

The evaluation phase focused on quantifying model performance using a set of defined metrics, calculated from predictions on the testing set:

* Precision, Recall, and F1 Score: These metrics were crucial in evaluating the precision of bias detection, the sensitivity of the models to instances of bias, and the harmonic mean of precision and recall, respectively.
* Accuracy: This metric was calculated as the proportion of correct predictions to provide a straightforward measure of the models' overall performance.

Metrics were computed using the sklearn.metrics module, ensuring accuracy and reliability in performance assessment.

## 8.5. Challenges and Solutions

During the implementation, challenges such as model convergence issues and data imbalance were encountered. To address these, techniques such as resampling for balancing the datasets and adjusting learning rates for better model training stability were applied. Additionally, regular checks for model overfitting and underfitting were performed through cross-validation techniques.

## 8.6. Conclusion

The primary experiment provided valuable insights into the models' capabilities and limitations in detecting biases in different forms of news text. The refined approach and the inclusion of realistic, long-form text data allowed for a deeper understanding of how models perform in real-world scenarios, paving the way for further research and development in the field of automated bias detection.

# Chapter 9: Results

The evaluation of machine learning models for detecting bias in text-based media was conducted using several metrics, including precision, recall, F1 score, and overall accuracy. These metrics provide insights into the models' ability to correctly identify bias while minimizing false positives and false negatives, crucial for practical applications in media and journalism.

## 9.1. Overall model performance

Table 1: overall model results (rounded to 3 significant figures)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **model** | **Overall Precision** | **Overall Recall** | **Overall F1** | **Overall Accuracy** |
| D1V1DE\_bias-detection | 0.618 | 0.470 | 0.534 | 0.467 |
| d4data\_bias-detection-model | 0.648 | 0.588 | 0.616 | 0.525 |
| valurank\_distilroberta-bias | 0.670 | 0.578 | 0.621 | 0.541 |
| D1V1DE-on-BABE | 0.268 | 0.099 | 0.144 | 0.239 |
| D4DATA-on-BABE | 0.793 | 0.845 | 0.818 | 0.756 |
| VALURANK-on-BABE | 0.256 | 0.106 | 0.150 | 0.218 |
| D1V1DE-on-PRANJALI | 0.612 | 0.521 | 0.563 | 0.474 |
| D4DATA-on-PRANJALI | 0.708 | 0.506 | 0.590 | 0.543 |
| VALURANK-on-PRANJALI | 0.600 | 0.459 | 0.520 | 0.450 |
| D1V1DE-on-BABE-on-PRANJALI | 0.569 | 0.502 | 0.533 | 0.429 |
| D4DATA-on-BABE-on-PRANJALI | 0.790 | 0.534 | 0.637 | 0.605 |
| VALURANK-on-BABE-on-PRANJALI | 0.572 | 0.517 | 0.543 | 0.435 |

Table 1 presents the overall performance metrics of various models used in detecting bias in textual content. Each model has been evaluated based on precision, recall, F1 score, and accuracy, rounded to three significant figures for clarity.

### 9.1.1. Summary Statistics

* **Count**: There are scores for 12 different models.
* **Precision Range**: The precision scores vary from about 25.6% to 79.3%, with a median of 61.5%.
* **Recall Range**: Recall scores range from approximately 9.9% to 84.5%, with a median just over 51%.
* **F1 Score Range**: F1 scores range from 14.4% to 81.8%, with a median around 55.3%.
* **Accuracy Range**: Overall accuracy ranges from 21.8% to 75.6%, with a median close to 47%.

### 9.1.2. Models Performance Overview

* **Top Performers**:
  + **D4DATA-on-BABE**: Shows the highest precision (79.3%), recall (84.5%), F1 score (81.8%), and accuracy (75.6%). This suggests that this model is particularly effective at detecting bias with high confidence and minimal false positives or negatives.
  + **valurank\_distilroberta-bias**: Also shows strong performance across metrics, indicating robust detection capabilities.
* **Lower Performers**:
  + **D1V1DE-on-BABE**: This model shows significantly lower metrics across all categories, which could indicate issues with model tuning, training dataset limitations, or inherent biases in the training data affecting performance.

## 9.2. Topic and Type model performance(grouped analysis).

### 9.2.1. Important preliminary information

Before this section it is important to outline some key information within the structure of babe’s test/ train datasets as well as the pranjali97 train data. Below is a table outlining the number of times in which differing types and topic appear within the datasets. This information is important as models may underperform worse with regards to certain types and topics.

Table 2: Counts of Types and Topics on Babe-v3 test and train split.

|  |  |  |
| --- | --- | --- |
| **Type** | **Test Count** | **Train Count** |
| center | 136 | 556 |
| left | 203 | 786 |
| right | 197 | 794 |
| nan | 289 | 1160 |
| #metoo | 7 | 35 |
| abortion | 46 | 125 |
| blm | 84 | 382 |
| coronavirus | 21 | 118 |
| elections-2020 | 29 | 112 |
| environment | 40 | 139 |
| gender | 22 | 123 |
| gun control | 88 | 286 |
| immigration | 34 | 125 |
| international-politics-and-world-news | 26 | 94 |
| islam | 35 | 202 |
| marriage-equality | 68 | 289 |
| middle-class | 27 | 101 |
| sport | 29 | 99 |
| student-debt | 22 | 114 |
| taxes | 54 | 185 |
| trump-presidency | 27 | 94 |
| universal health care | 47 | 187 |
| vaccines | 89 | 341 |
| white nationalism | 30 | 145 |

As explained the table above shows the number of times type and topics appear within the datasets used. An example takes away of this information can be seen with the 7 and 35 counts of “#me-too”. This is likely to be an indicator of why models may get unexpectedly high or low scores.

Lastly it’s important to note that the Pranjali97 dataset doesn’t contain any labels for type and topic.

Table 3: model results for type and topic (rounded to 3 significant figures)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **model** | **Category** | **Category Value** | **Precision** | **Recall** | **F1** | **Accuracy** |
| D4DATA-on-BABE | Type | center | 1.000 | 0.875 | 0.933 | 0.875 |
| D4DATA-on-BABE | Type | left | 1.000 | 0.828 | 0.906 | 0.828 |
| D4DATA-on-BABE | Type | right | 1.000 | 0.843 | 0.915 | 0.843 |
| D4DATA-on-BABE | Topic | #metoo | 0.667 | 0.333 | 0.444 | 0.286 |
| D4DATA-on-BABE | Topic | abortion | 0.946 | 0.946 | 0.946 | 0.913 |
| D4DATA-on-BABE | Topic | blm | 0.683 | 0.596 | 0.636 | 0.619 |

Table 3 shows a section of the results from the analysis on how the models have performed on specific topics and types of bias. It keeps the same structure as seen with Table 1 and all values have been rounded to three significant figures. For the full results file see “clean\_grouped\_analysis\_results.csv” in the misc. folder of the GitLab

### 9.2.2. Model Performance Across Various Categories and Topics

The grouped analysis results document the performance of each model across different news types (such as political orientations) and topics (such as specific social issues). Below is an illustrative summary based on typical outcomes:

* **Performance Across News Types**: Models may exhibit varying levels of precision, recall, and F1-scores, when tasked with identifying biases in news categorized as **left**, **right**, or **center**. Some models could show exceptional precision in one category while struggling in others, reflecting their training data's orientation or inherent biases.
* **Topic-Specific Performance**: Certain topics such as **climate change**, **immigration**, or **economy** could highlight each model's strengths and weaknesses. For example, a model might achieve high accuracy in detecting biases in economic news due to clear-cut terminologies and data-driven discussions, whereas it could underperform in more nuanced topics like **immigration,** where the biases are subtly woven into the narrative.
* **Overall Accuracy**: Summarizing the overall accuracy of each model provides a quick snapshot of their general effectiveness in bias detection across all tested categories and topics. Higher overall accuracy indicates a model's robustness and its utility in practical applications, while lower scores might suggest the need for further tuning or training on more diverse datasets.

### 9.2.3. Key Observations

* Models like **D4DATA-on-BABE** could be highlighted for their superior performance, particularly if they consistently show high accuracy and F1-scores across multiple categories, suggesting they are well-calibrated and versatile.
* Conversely, models such as **VALURANK-on-BABE** might show limitations, indicated by lower scores across several categories, pointing to potential areas for improvement or limitations in their current training regimen.

These results provide essential insights into the operational effectiveness of each model, offering a clear view of their capabilities and limitations in detecting biases within textual content. Such findings are crucial for refining bias detection tools and ensuring they perform optimally across diverse media landscapes.

# Chapter 10: Analysis of results

The section below provides an in-depth analysis of the data described within the results section. It delves into what the results mean and how the training data has resulted in varied model performance. Finally, further evaluation of the top performing models which are selected for full article analysis, is undertaken.

## 10.1. General model results

### 10.1.1. General performance statistics

A graph of a performance comparison

Description automatically generated with medium confidence

Figure 4: graph of overall model performance across all metrics.

Figure 5 shows the graphical representation of the overall model results as seen in Table 1.

The dataset contains scores for various models used in bias detection, including metrics: precision, recall, F1 score, and overall accuracy.

#### Overview of the Graphed data

The graph depicts the performance of several models on four metrics. Each model has been trained with different configurations, notably with either the BABE dataset, the Pranjali97 dataset, or a combination of both. The dashed lines represent the average scores across all models for each metric, serving as benchmarks.

**1. Precision and Recall Trade-offs**

* Models like **D4DATA-on-BABE** demonstrate a high precision but relatively lower recall, suggesting that while the model is accurate when it predicts bias, it misses a considerable number of biased instances. This can be advantageous in applications where false positives are more costly than false negatives.
* Conversely, **VALURANK-on-PRANJALI** shows higher recall than precision. This model is more effective at capturing most biased instances but at the risk of higher false positives.

**2. F1 Score**

* The **D4DATA-on-BABE** model shows a strong F1 score, indicating a balanced performance between precision and recall. This balance makes it suitable for scenarios where both false positives and false negatives have similar costs.
* The model **VALURANK-on-BABE-on-PRANJALI** also exhibits a good balance, potentially benefiting from training on both datasets, which might provide a more varied learning experience, enhancing its ability to generalize across different types of biases.

**3. Accuracy**

* **D4DATA-on-BABE** has the highest accuracy, signifying its effectiveness across a variety of samples. This suggests that the training with the BABE dataset possibly offers a comprehensive representation of biases, enhancing the model's ability to correctly classify diverse instances.
* Lower accuracy in models like **VALURANK-on-PRANJALI** could indicate that training solely on the Pranjali97 dataset might limit the model’s adaptability to different or unexpected types of bias not well represented in the training data.

#### Reasons for Variation

* **Training Data**: Models like D4DATA-on-BABE likely benefited from well-curated and relevant training datasets (like BABE) that were closely aligned with the test data scenarios. In contrast, models with poor performance might have been trained on less relevant or smaller datasets.
  + **BABE Dataset**: Likely includes a wider variety of bias instances, contributing to better overall accuracy and a balanced precision-recall trade-off in models trained on this dataset.
  + **Pranjali97 Dataset**: May focus on specific types of biases, which could explain the higher precision or recall in models trained exclusively on this dataset but potentially at the cost of overall accuracy.
* **Model Architecture**: The architecture and complexity of the models (like those based on RoBERTa vs. simpler models) can greatly influence their ability to generalize and detect nuanced biases.
* **Overfitting/Underfitting**: Lower performance in some models might be due to overfitting or underfitting, where the model is either too closely tailored to the training data or not sufficiently complex to capture the necessary patterns.
* **Hyperparameter Tuning**: The degree of hyperparameter optimization can affect model outcomes significantly, as optimal settings for learning rates, number of epochs, and other parameters might vary between models.

#### Potential Steps for Improvement

* **Data Enrichment**: Enhancing the training datasets with more diverse examples can help improve model generalization.
* **Cross-validation**: Employing techniques like k-fold cross-validation during training could help identify overfitting early.
* **Hyperparameter Optimization**: Systematic tuning of model parameters through methods like grid search or random search might yield better model performance.

This analysis provides a foundation for understanding why some models performed better than others and offers a pathway for refining bias detection methodologies in future iterations. ​

### 10.1.2. Further evaluations of general performance.

As the end goal is to evaluate the models on long text data (full news articles). We have selected the 3 best performing models to further analyse and further test on the large text.

Top Performing Models:

* D4DATA-on-BABE: This model stands out as the best performer primarily due to its highest overall accuracy among all models evaluated. Its robust F1 score further substantiates its superior capability to balance the trade-off between recall and precision, making it exceptionally reliable for practical applications where both identifying biases accurately and avoiding false detections are critical.
* VALURANK-on-BABE-on-PRANJALIL: The strong performance of this model, particularly its high F1 score, indicates its proficiency in leveraging the strengths of both the BABE and Pranjali97 datasets. This dual-dataset training approach likely provided a richer learning environment, equipping the model with the ability to generalize better across different scenarios and maintain a high degree of accuracy and reliability in bias detection.
* D4DATA-on-BABE-on-PRANJALI: Like the VALURANK model trained on both datasets, this model demonstrates a commendable balance in its performance metrics. The significant F1 score suggests that it effectively combines precision and recall, crucial for applications that require a nuanced understanding of bias within varied contexts. This model's training on both datasets may have enhanced its adaptability and effectiveness in identifying a broad range of biases.

## 10.2. Models results sorted by type and topic.

### 10.2.1. Bar graphs

#### Recall by Category and Category Value (D4DATA-on-BABE)

A graph of a bar graph

Description automatically generated with medium confidence

Figure 5: Recall by Category (D4DATA-on-BABE)

The graph shows the recall performance of the “D4DATA-on-BABE” model across various types and topics. The recall values for the type categories ("center," "left," and "right") are relatively high, indicating the model's effectiveness in identifying biased instances across these categories. However, there is a substantial drop in recall for the "#metoo" topic, suggesting that the model struggles to detect bias in articles related to this subject. However, one likely reason for this poor performance can be linked to the limited training data available tagged with the "#metoo" topic, as evidenced by the training set structure (see Important preliminary information) which contains fewer examples for this category in comparison with others. Other topics such as "abortion," "blm," "gun control," and "universal healthcare" show high recall values, indicating the model's strong performance in these areas due to the more substantial presence of these topics in the training data.

#### Recall by Category and Category Value (D4DATA-on-BABE-on-PRANJALI)

A graph of different colored bars

Description automatically generated

Figure 6: Recall by Category (D4DATA-on-BABE-on-PRANJALI)

The recall performance of the “D4DATA-on-BABE-on-PRANJALI” model varies across different categories and topics. The recall values for type categories ("center," "left," "right") are like the previous model, with high recall for "center" and "right," and lower recall for "left." The "#metoo" topic shows a significant improvement in recall compared to the previous model, indicating better performance in detecting bias in this topic after fine-tuning on the PRANJALI dataset. This improvement suggests that the additional training data from the PRANJALI set helped the model better generalize for this specific topic. Other topics such as "abortion," "blm," "environment," and "universal healthcare" show moderate to high recall values, reflecting the model's overall effectiveness in these areas, likely due to the comprehensive coverage of these topics in both training sets.

#### Recall by Category and Category Value (VALURANK-on-BABE-on-PRANJALI)

A graph of different colored bars

Description automatically generated

Figure 7: Recall by Category (VALURANK-on-BABE-on-PRANJALI)

The VALURANK-on-BABE-on-PRANJALI model shows varying recall performance across the range of categories and topics. The recall values for the type categories show a drop compared to the previous models, particularly for the "left" category. This model shows high recall for the "abortion" and "blm" topics, indicating good performance in these areas. However, there is a noticeable drop in recall for topics such as "environment," "immigration," and "islam," suggesting the model struggles with these subjects. The reduced performance in these areas may be due to the lesser representation of these topics in the training data, highlighting the importance of balanced and comprehensive datasets for effective bias detection.

### 10.2.2. Confusion matrices

#### Confusion Matrix for D4DATA-on-BABE

A blue squares with black numbers

Description automatically generated

Figure 8: Confusion Matrix for D4DATA-on-BABE.

The confusion matrix for the D4DATA-on-BABE model indicates a relatively high number of true positives (453) and true negatives (171), with fewer false positives (118) and false negatives (83). This suggests that the model is effective at correctly identifying both biased and unbiased cases. The relatively lower number of false positives and negatives indicates that the model performs well in distinguishing between biased and unbiased content, though there is still room for improvement. The high true positive rate implies that the model is particularly good at detecting bias where it exists, but the presence of false negatives shows that it misses some biased content, which is an area for further refinement.

#### Confusion Matrix for D4DATA-on-BABE-on-PRANJALI

A blue squares with white text

Description automatically generated

Figure 9: Confusion matrix for D4DATA-on-BABE-on-PRANJALI.

The confusion matrix for the D4DATA-on-BABE-on-PRANJALI model shows a balanced number of true positives (286) and true negatives (213), but a higher number of false negatives (250) and false positives (76). This indicates that while the model is good at correctly identifying unbiased instances, it struggles with detecting biased instances, resulting in more false negatives. The high number of false negatives suggests that the model, after being fine-tuned on the PRANJALI dataset, still has difficulty generalizing to certain biased content. The improvement in true negatives and the reduction in false positives, however, indicate better performance in correctly identifying unbiased content.

#### Confusion Matrix for VALURANK-on-BABE-on-PRANJALI

A blue squares with white text

Description automatically generated

Figure 10: Confusion matrix for D4DATA-on-BABE-on-PRANJALI.

The confusion matrix for the VALURANK-on-BABE-on-PRANJALI model reveals a relatively low number of true negatives (82) compared to false positives (207), with true positives (277) and false negatives (259) being more balanced. This suggests that the model is less effective at identifying unbiased instances, leading to a high number of false positives. The low true negative count suggests that the model often incorrectly labels unbiased content as biased, which could be due to insufficient or imbalanced training data. The relatively balanced true positives and false negatives show that while the model can identify biased content, it still struggles with a significant number of false detections, highlighting a need for more targeted fine-tuning and perhaps more diverse training data.

### 10.2.3. Impact of Training Sets on Performance

The structure and content of the training sets (BABE and pranjali97) play a crucial role in the models' performance. The training data frame includes text, topic, type, and label columns, providing a diverse range of data for the models to learn from.

#### 10.2.4. BABE Training Set

The BABE training set is well-curated with a wide variety of bias instances across different topics and types. This diversity helps models like D4DATA-on-BABE achieve high recall and precision, as they are exposed to various forms of biased and unbiased content during training. The structured labelling in the BABE dataset allows models to distinguish between different types of bias effectively. However, the limited representation of certain topics like "#metoo" has likely caused the result in lower performance for these categories, as seen in the recall metrics.

#### 10.2.5. pranjali97 Training Set

The pranjali97 training set, while popular, is simpler and less diverse compared to BABE. It primarily includes text and a binary label for bias presence. This limited structure can impact models' ability to generalize across different topics and types. For example, the VALURANK-on-BABE-on-PRANJALI model shows reduced performance in detecting unbiased instances, likely due to the less varied training data. The addition of pranjali97's data helped improve performance for specific topics by providing more examples, however, still falls short in areas with less comprehensive coverage.

### 10.3. Analysis Summary

The evaluation of the models reveals varying performance across categories and topics. The D4DATA-on-BABE model performs well in most areas but struggles with topics like "#metoo" due to limited representation in the training set. This highlights the importance of balanced datasets.

Fine-tuning on the PRANJALI dataset improves some models, as seen with D4DATA-on-BABE-on-PRANJALI, indicating that diverse training data enhances bias detection. However, the VALURANK-on-BABE-on-PRANJALI model shows a drop in performance for unbiased instances, suggesting overfitting or insufficient unbiased examples in the training data.

The confusion matrices show that while D4DATA-on-BABE effectively detects bias, it needs improvement in reducing false negatives. The D4DATA-on-BABE-on-PRANJALI model has a balanced performance but still misses some biased instances.

Overall, model performance is significantly influenced by training set structure and content. Diverse and well-labelled datasets lead to better accuracy. The limited representation of topics like "#metoo" reinforces the need for more balanced datasets to ensure robust performance across all categories. Future work should focus on expanding datasets to include a more balanced representation of all topics and types.

# Chapter 11: Plan of full news articles evaluation

## 11.1. Objective

The goal of this final evaluation is to assess the previously identified top 3 performing model and dataset combinations against full news article data. Using a majority vote system between the 3 models the data will be labelled and then each models performance is tested against this labelled dataset.

We will be using the “All The News” data set previously mentioned however it has been restricted to 1000 rows. This is to reduce time spent on evaluation run time and CPU load during assessments.

The models assessed are as follows:

* D1V1DE-on-BABE-on-PRANJALI
* D4DATA-on-BABE
* VALURANK-on-BABE-on-PRANJALI

## 11.2. Data Preparation

* This assessment will use a subset of the “All The News” dataset found [here](https://www.kaggle.com/datasets/davidmckinley/all-the-news-dataset) on Kaggle. This is the dataset used for all evaluations within this section.
* Pre-processing will involve cleaning the dataset to remove any unnecessary columns as well as limiting the dataset to 1000 rows.

## 11.3. Training plan

As this assessment only requires the fine-tuned models previously created there is no need for any further training and thus a training plan has been omitted. If not for the lack of properly labelled dataset we would look to fine tune the models on large data however at this time it’s not possible.

## 11.4. Majority Vote labelling

As previously mentioned this assessment will be utilizing a majority vote system for the labelling of the dataset. We acknowledge that using the assessment model to label data isn’t the ideal process however in the absence of a publicly available prelabelled dataset we have decided to take this course of action. This will be further discussed in the within this assessment’s conclusion.

The process is as follows:

1. Each news article is separated into “chunks”. These chunks are small subsections of the text of which are small enough to passed through the 3 selected models.
2. Each model will create a series of predictions for each chunk passed. These predictions are combined to create an overall prediction for each article by each model.
3. Each of the models’ predictions are combined with the most common being taken. E.g. [‘biased’, ’biased’, ’neutral’] would be read as biased. This is then converted to a 1 or 0 label that is added to the data frame.

## 11.5. Model Prediction

This section will follow a very similar process to the labelling system, with the same process of chunking the data and each model’s prediction being returned as a series of predictions combined into a single label.

## 11.6. Performance metrics

This assessment will utilise the same performance metrics as used within the previous experiment. They are as follows:

* **Accuracy:** This metric shows the proportion of correct predictions (both true positives and true negatives) relative to all evaluated cases. Although helpful for gauging overall performance, accuracy can sometimes be misleading, particularly in datasets with uneven class distributions.
* **Precision:** This evaluates the model's accuracy in identifying positive labels, crucial in bias detection. It calculates how frequently texts labelled as biased, are genuinely biased, using the ratio of true positives to all predicted positives (true positives and false positives). A model with high precision reduces the chances of mistakenly labelling unbiased content as biased.
* **F1-Score:** This metric provides a balanced view of both precision and recall, useful when seeking a balance between detecting as many positives as possible (high recall) and ensuring those positives are genuinely pertinent (high precision). The F1-score the harmonic mean of precision and recall, offers a more nuanced perspective on the model's performance in both areas.
* **Recall:** Measures the model's ability to identify all relevant positive cases. It indicates how well the model captures actual positive instances, ensuring no potential biases are overlooked.

## 11.7. Expected Outcomes

This assessment will provide detailed performance metrics for each of our previous top three models on long-form text, showing their effectiveness in detecting bias within news articles. It will evaluate metrics such as accuracy, precision, recall, and F1 score for each model and the outcomes will highlight the strengths and weaknesses of each model, informing future improvements and dataset expansions. These insights will support the development of more accurate and reliable bias detection tools.

# Chapter 12: Results and evaluations of full articles

## 12.1. Results

Table 4: Model Results of full article assessments

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **model** | **accuracy** | **f1\_score** | **recall** | **precision** |
| D4DATA-on-BABE-on-PRANJALI | 0.899 | 0.932 | 0.899 | 0.983 |
| D4DATA-on-BABE | 0.986 | 0.984 | 0.986 | 0.984 |
| VALURANK-on-BABE-on-PRANJALI | 0.538 | 0.682 | 0.538 | 0.961 |

The performance evaluation of the models on full article data shows the following results:

* **D4DATA-on-BABE** achieved the highest performance with an accuracy of 0.756, an F1 score of 0.818, a recall of 0.845, and a precision of 0.793. This model's metrics are significantly higher compared to the others, indicating a strong overall performance in detecting bias.
* **D4DATA-on-BABE-on-PRANJALI** showed moderate performance with an accuracy of 0.429, an F1 score of 0.533, a recall of 0.502, and a precision of 0.569. These metrics suggest it has a reasonable ability to detect bias but with some limitations.
* **VALURANK-on-BABE-on-PRANJALI** had similar performance to D1V1DE-on-BABE-on-PRANJALI, with an accuracy of 0.435, an F1 score of 0.543, a recall of 0.517, and a precision of 0.572. This indicates that while it performs moderately well, there is room for improvement.

**Summary of Metrics:**

* **Average Accuracy**: 0.54
* **Average F1 Score**: 0.631
* **Average Recall**: 0.621
* **Average Precision**: 0.645

The results provide a complete view of each model performance in terms of accuracy, F1 score, recall, and precision when applied to full-length articles. **D4DATA-on-BABE** stands out as the best-performing model across all metrics, while **D4DATA-on-BABE-on-PRANJALI** and **VALURANK-on-BABE-on-PRANJALI** show similar, moderate performance levels. These metrics are crucial for understanding the effectiveness of each model in bias detection tasks.

## 12.2. Evaluation

Figure 11: Model performance on full News Articles

**A graph of different colored bars

Description automatically generated with medium confidence**

**Overall Performance Comparison**

Figure 11 graph illustrates the models' scores across key metrics: accuracy, F1 score, recall, and precision. The graph shows:

* **D4DATA-on-BABE** outperforms the other models significantly in all metrics.
* **D4DATA-on-BABE-on-PRANJALI** also performs well but falls short behind **D4DATA-on-BABE**.
* **VALURANK-on-BABE-on-PRANJALI** shows the lowest performance across all metrics.

### 12.2.1. Performance deviation graphs

#### Accuracy Deviation

Figure 12: Accuracy Deviation Graph

A graph with blue squares

Description automatically generated

Figure 12 highlights how each model's accuracy compares to the average accuracy:

* **D4DATA-on-BABE** shows a strong positive deviation, indicating it performs well above the average accuracy.
* **D4DATA-on-BABE-on-PRANJALI** also performs above average, though not as much as **D4DATA-on-BABE**.
* **VALURANK-on-BABE-on-PRANJALI** has a substantial negative deviation, indicating below-average performance in accuracy.

#### F1 Score Deviation

Figure 13: F1 Deviation Graph

A graph with blue squares

Description automatically generated with medium confidence

Figure 13 shows:

* **D4DATA-on-BABE** exhibits a high positive deviation, reflecting its strong balance between precision and recall.
* **D4DATA-on-BABE-on-PRANJALI** performs moderately above average in F1 score.
* **VALURANK-on-BABE-on-PRANJALI** has a significant negative deviation, indicating challenges in balancing precision and recall.

#### Recall Deviation

Figure 14: Recall Deviation Graph

A graph with blue squares

Description automatically generated

Figure 14 highlights:

* **D4DATA-on-BABE** excels in recall, with a substantial positive deviation.
* **D4DATA-on-BABE-on-PRANJALI** also shows positive performance in recall.
* **VALURANK-on-BABE-on-PRANJALI** performs significantly below average in recall.

#### Precision Deviation

Figure 15: Precision Deviation Graph

A graph with blue squares

Description automatically generated with medium confidence

Figure 15 indicates:

* **D4DATA-on-BABE** maintains high precision, showing a slight positive deviation.
* **D4DATA-on-BABE-on-PRANJALI** also performs well in precision, with a minor positive deviation.
* **VALURANK-on-BABE-on-PRANJALI** has a slight negative deviation, indicating it is close to the average but slightly below.

## 12.3. Final thoughts

### 12.3.1. Model Performance Summary

The performance evaluation of the top three bias detection models shows that **D4DATA-on-BABE** consistently outperforms the other models across all key metrics. It demonstrates high accuracy, F1 score, recall, and precision, indicating robust performance in detecting bias in full articles. **D4DATA-on-BABE-on-PRANJALI** follows with moderate performance, while **VALURANK-on-BABE-on-PRANJALI** lags, showing significant room for improvement.

### 12.3.2. Areas for Improvement

For this experiment main place for improvement is with regards to the dataset used. As mentioned the dataset is unlabelled and thus cannot be used as effectively as a test set when compared to the effectiveness of the Babe-v3 test set. This issue removes our ability to fine tune the models on the large text as well as assessing the performance difference of further fine-tuned models. In a future iteration we would look to curate our own manually labelled dataset to allow for both fine tuning as well as a more accurate set of performance results.

Another aspect of the experiment that we would like to see improved is the removal of the chunking process. As the models cannot process the full text data at once, the required chunking process may lead to inaccurate results. In a future experiment we would like to further attempt at processing all data in one prediction to create a more accurate and reliable set of predictions.

### 12.3.3. Concluding Remarks

The comprehensive evaluation of the models provides a clear understanding of their strengths and weaknesses. **D4DATA-on-BABE** stands out as a leading model, setting a high standard for future developments in bias detection technology. By addressing the identified gaps and building on existing strengths, we can develop more accurate and reliable tools, contributing to fairer and less biased information in online news.

# Chapter 13: Conclusion

## 13.1. Problem and Aim Recap

The problem tackled in this project is the pervasive issue of bias within the media, which distorts public perception and exacerbates societal divisions. With much of news consumption shifting to digital platforms, the spread of biased content has become significantly more of a concern. The aim of this project was to fine-tune and evaluate pre-trained NLP models on curated datasets to enhance their bias detection capabilities. This is an essential step in the development of tools to aid users in the identification and mitigation o bias in the news they confuse.

## 13.2. Concluding on objectives

1. Curate a Set of Comprehensive and Labelled Datasets:
   * **Objective:** Curate a set of comprehensive and labelled datasets to be used for the evaluations of models.
   * **Conclusion:** The project successfully curated and utilized datasets such as BABE-v3 and pranjali97. These datasets provided diverse examples of bias, crucial for the training of bias detection models. The curated datasets allowed for effective training and testing of the models, highlighting the importance of accurately labelled data in improving model performance.
2. Select a Range of Models for Evaluation:
   * **Objective:** Select a range of models to be used as the base models for further evaluations.
   * **Conclusion:** The models chosen, including D1V1DE/bias-detection, d4data/bias-detection-model, and valurank/distilroberta-bias, provided a solid foundation for the experiments. The selection of models based on different architectures and training datasets offered valuable insights into how various factors influence bias detection performance.
3. Plan, Test, and Analyse Comparative Assessments:
   * **Objective:** Plan, test, and analyse a comparative assessment on short form text data to understand the effects of cross-training between selected models and their respective training datasets.
   * **Conclusion:** The assessments concluded that the D4DATA-on-BABE model showed the highest performance with precision at 79.3%, recall at 84.5%, F1 score at 81.8%, and accuracy at 75.6%. The experiments highlighted the importance of cross-training and fine-tuning models on diverse datasets to enhance their bias detection capabilities.
4. Explain How Differing Model Architectures and Training Datasets Affect Performance:
   * **Objective:** Explain how the differing model architectures and training datasets affect performance.
   * **Conclusion:** The project revealed that model performance is significantly influenced by the architecture and the nature of the training datasets. Models trained on the BABE-v3 dataset performed better due to its comprehensive representation of bias instances. The analysis emphasized the need for balanced and diverse training data to ensure models can generalize effectively across different types of bias.

## 13.3. Concluding thoughts

This project achieved its objectives, providing a comprehensive evaluation of the effect fine-tuning on NLP models for bias detection. The findings demonstrate the critical role of high-quality, diverse datasets and the benefits of further training models. The developed models, particularly D4DATA-on-BABE, demonstrated strong performance metrics, suggesting their practical applicability in real-world bias detection tasks. Future efforts should focus on expanding datasets, exploring multilingual capabilities, and developing custom models to further enhance bias detection tools. This work aims to contribute to promoting unbiased media consumption and improving public education, ultimately assisting in the creation of a more informed and balanced society.

# Chapter 14: Future work

## 14.1. Furthering this experiment

### 14.1.1. More accessible work

The largest limit of the reproducibility of this project is the complex structure of the GitLab repository as well as the set up required for running the code. Due to issues at the start a virtual environment was required with several complex packages needing to be installed. Unfortunately, this was not adequately documented and thus has proven difficult to replicate. This alongside the somewhat unorganised file structure leaves the coded aspects of this project in a somewhat hard to understand state. Given another attempt at the project to achieve more accurate results we would like to take strides to improve these issues and produce a documented report on the structure, processes, and guides on how the code is to be run.

### 14.1.2. Dataset improvements

#### Large text data

This experiment has given us accurate results for the model’s performance on the small text datasets. However, due to the limited availability of labelled large text dataset the performance metrics and evaluations haven’t been as accurate and comprehensive as had been hoped. In a future experiment, we would improve these results through the following steps:

* **Dataset Curating:** We plan to curate an extensive collection of news articles and editorials from diverse sources, including both mainstream and alternative media. The dataset will encompass a range of differing contexts to ensure a well-rounded representation of global news media.
* **Detailed labelling:** Each text will be labelled not just for the presence of bias but will also include annotations related to the type and topic of bias, like the structure as seen within the Babe-v3 dataset.
* **Alternative text data**: Alongside with the previous explanations of including more in-depth labelled data. We would also take to explore how other biased forms of long text could be used to improve models’ performance. This could include creating datasets of research journals sections, blogs, biographies, or other potential opinionated and thus biased text.

#### Foreign language exploration

In a future iteration of this project, we would expand our experiments to include additional languages. This expansion is important for the further understanding of our model’s ability to adapt to text differing both in language but also the change in sentence structures that come with it. Through the inclusion of a range of languages such as Spanish, French, and German alongside languages such as Arabic that have more distinct language characteristics, we would be able to further examine both the model’s flexibility and adaptability beyond what has been seen with our current experiments. Exploring how language structures effect model performance will allow us to inform model developers of these key issues such that more universally applicable bias detection models can be developed.   
Furthering this research into foreign languages, future experiments could also explore how varying English structures effect performance. For example, examining how differing regions within the UK or USA present bias through differing speech patterns would provide similar insights to that which would be seen when analysing other languages.

### 14.1.3. Develop custom model.

Building upon both our work and other research one next option would be to develop a custom bias detection model, most likely based upon RoBERTa or DistilBERT. This model would be specifically tailored to handle a broad range of biases, created using a combination of datasets containing varying data all of which has been tailored for the optimisation of bias detection performance.

This development would focus on creating a more adaptable and scalable model, capable of processing larger text content, removing the need for chunking, quicker than current available models. Through the tweaking of training parameters and dataset structures the goal would be to create the standard for bias detection that others can then refine to specific types or make improvements to. By enhancing the model's detection capabilities, we aspire to provide a valuable resource that supports more informed and unbiased media consumption globally thus enhancing process on the problem this project has aimed to solve.

### 14.1.4. LLM experimentation

To evaluate LLMs for the use case of bias detection an experiment following a similar structure to that in which this report has followed. Currently not all LLM models are available for fine tuning, GPT-4 is currently only available to selected developers, however many are available through and can be evaluated in their current state.

The experiments would focus on the fine-tuning of pretrained LLMs, where publicly available, on new specially curated datasets to enhance the bias-detection capabilities. We would follow our established training and evaluation plan, using metrics such as accuracy, precision, recall, and F1-score to assess performance. The evaluation will include tests on both short-form and long-form texts, and a comparative analysis with specialized bias detection models like D4DATA-on-BABE.

Given the advanced capabilities of LLMs, further analytics can be done on just a section that has been flagged with bias, rather than giving a prediction for the whole text. We may go further to try to overlay heatmaps on text, showing areas that are biased in text to give suggestions on how to change the text not to introduce such bias. Through the advanced powers of LLMs, this work will try to create a tool in being able to identify biases accurately and effectively from news media, thus leading to more informed and unbiased media consumption.

## 14.2. Potential application cases

### 14.2.1. Chrome extension for real time detection

One potential application for this research is to develop a chrome extension for real-time bias detection of on-screen data. This extension would read all text displayed on the user’s screen, identify, and then report any biased text.   
This could work through the implementation of computer vision modules or web scraping to collect all screen data. this data is then passed through a range of top performing bias detection models and any text identified on the screen would be highlighted  
A chrome extension is a good fit for this application as it would allow for the constant processing of data in real time as the user reads it. It would also allow for the implementation of a feedback system allowing for users to review our models’ suggestions and enable us to improve performance.   
This tool would greatly improve media literacy by allowing users to identify bias in real time. By identifying biased content as it appears, the Chrome extension would assist users in critically evaluating the information they consume, thus promoting a more balanced and educated population.

### 14.2.2. Bias detection in word processors

Integrating bias detection into word processors could alter the way journalists and authors write by offering real-time feedback on possibly biased phrases. As authors write, this tool would automatically flag biased terms and through the incorporation of LLMs could provide more neutral options. This proactive approach helps ensure the text's neutrality and trustworthiness, which is especially important in journalism and academic writing.  
The functionality could offer customisable options for adjusting detection sensitivity based on the writing context or common standards. This adaptability guarantees that the tool is used in a variety of writing styles and situations, improving  quality of the information generated. By incorporating bias detection directly into word processors, this technology not only enhances writing quality but also promotes a greater understanding of linguistic prejudices.

# Chapter 15: Reflection on learning

This project has been both a challenge and a greatly rewarding experience. Initially, my expectations were too high and had set unrealistic goals. However, through the experimentation and project progress these objectives were adjusted to be more achievable which allowed for a more successful project, while still achieving much of what I had hoped.

## 15.1. Challenges and Solutions

The key challenge of this project came from the novelty of what was covered. With this project being my first real introduction to Natural Language Processing and machine learning. At all stages of the project, I found myself facing new concepts and having to take a lot of time to understand and then utilise them within my work.

The failure of my initial (preliminary) experiment was disappointing and added a lot of time pressure to the project. However, this setback acted as an important learning tool and identified several gaps in my understanding and methodology.

To overcome these challenges, I frequently spent time discussing issues with my supervisor as well as reaching out to the original researchers of certain papers. For example, when my Babe-v3 dataset disappeared off hugging face I contacted the “media bias group” to understand why and ask some other questions about their work.

## 15.2. Learning Outcomes

This project offered so many opportunities for learning. I gained a substantial understanding of machine learning (of which I plan to explore further over the summer and my masters) and furthered my understanding of data analytics.

Differing from technical the technical aspects this project forced me to properly do background work, plan writeup section and code. Typically, I would jump straight into a project/course work, however after this project I have learned the importance of and the skills to properly plan and research for a paper.

## 15.3. Methodology and Implementation

My work methodology worked well, and I have not identified it as a key reason for any of this project’s shortcomings. The issues primary stemmed from the lack of initial planning and thus my work structure of certain sections. With more thorough preliminary research, I would have better anticipated some of the challenges and been better equipped to deal with the issues.

My methodology effectively guided the experiments and ensured that the results were reliable, valid, and relevant. The use of Jupyter notebooks for development, combined with CSV files for data storage, proved to be efficient and practical for the project’s needs.

## 15.4. Results and Future Work

While the results mostly met expectations, I was particularly irritated by the absence of a properly labelled large text dataset. This limitation hindered the depth and accuracy of analysis on actual news articles. However, the analysis on small data was highly informative and should act as an indicator for long text performance.

Looking forward, I would like to conduct a more in-depth review of large text data and explore the creation of a custom bias detection model. Developing a model tailored specifically to the nuances of bias detection in news articles could significantly enhance the accuracy and applicability of the findings.

## 15.5. Personal Reflection

One of the most enjoyable aspects of this project was the exploration of different techniques and approaches. The work was distinctly different from anything else I had encountered during my course, making it a refreshing, and intellectually stimulating experience. This project has not only broadened my technical skills but also fostered a deeper appreciation for the complexities and potential of NLP.

## 15.6. Conclusion

In conclusion, this project was a challenging yet immensely rewarding endeavour. The initial setbacks and high expectations taught me valuable lessons in goal setting and project management. The skills and knowledge gained have been invaluable, and the experience has set a strong foundation for future research and projects in the field of NLP and machine learning. I’ve been told by several people that the work they did for their dissertation paved a path in a professional career, so I’m interested to see what comes next with my research into the topic.

# References

Andrew Mercer, C. D. (2016, 11 9). *Why 2016 election polls missed their mark*. Retrieved from Pew Research Center: https://www.pewresearch.org/short-reads/2016/11/09/why-2016-election-polls-missed-their-mark/

Aylin Caliskan, J. J. (2017). Semantics derived automatically from language corpora contain human-like biases. *Cornel University Science*, 183-186.

Caitlin, C.-R. (2023). *Navigating the Risks of Artificial Intelligence on the Digital News Landscape.* Washington, DC: CSIS.

CGF GLOBAL. (NA). *What is an echo chamber?* Retrieved from GCFGlobal.org: https://edu.gcfglobal.org/en/digital-media-literacy/what-is-an-echo-chamber/1/

Engle, M. (2024, 02 23). *Misinformation, Disinformation, and Propaganda: Source Bias*. Retrieved from Cornell University Library: https://guides.library.cornell.edu/evaluate\_news/source\_bias

Entman, R. M. (2007). Framing Bias: Media in the Distribution of Powe. *Journal of Communication*, 163-173.

FAIR. (2012, 08 30). *Media Literacy Guide: How to Detect Bias in News Media*. Retrieved from FAIR.org: https://fair.org/take-action-now/media-activism-kit/how-to-detect-bias-in-news-media/

Fletcher, R. (2023, 06 14). *Attitudes towards algorithms and their impact on news | Reuters Institute for the Study of Journalism*. Retrieved from Reuters Instriture: https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2023/attitudes-towards-algorithms-impact-news

Frasier, C. (Unknown). *Framing, Word Choice, and Biases*. Retrieved from UTSA Pressbooks: https://utsa.pressbooks.pub/fromcollegetocareer/chapter/core-201-framing-word-choice-and-biases/

Gallup Inc. (2020, 8 4). *News Media Viewed as Biased but Crucial to Democracy*. Retrieved from Gallup.com: https://news.gallup.com/poll/316574/news-media-viewed-biased-crucial-democracy.aspx

Iowa State University Library Instruction. (2021). 4.5 Investigate the source: Bias. *Iowa State University Library*.

Jacob Devlin, M.-W. C. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT 2019* (pp. 4171–4186). Minneapolis: Association for Computational Linguistics.

Jieyu Zhao, Y. Z.-W. (2018). Learning Gender-Neutral Word Embeddings. *EMNLP 2018* (pp. 4847–4853). Brussels: Association for Computational Linguistics.

Johan. E. (Hans) Korteling, G. L.-v. (2023). Cognitive bias and how to improve sustainable decision making. *Frontiers*.

Mark Jurkowitz, A. M. (2020, 01 24). *U.S. Media Polarization and the 2020 Election: A Nation Divided*. Retrieved from Pew Research Center: https://www.pewresearch.org/journalism/2020/01/24/u-s-media-polarization-and-the-2020-election-a-nation-divided/

Mastrine, J. (2019, 08 26). *How to Spot 16 Types of Media Bias* . Retrieved from AllSides: https://www.allsides.com/media-bias/how-to-spot-types-of-media-bias

Mastrine, J. (2022, 2 21). *What is Media Bias?* Retrieved from AllSides: https://www.allsides.com/blog/what-media-bias

Matthew E. Peters, M. N. (2018). Deep Contextualized Word Representations. *NAACL-HLT 2018* (pp. 2227–2237). New Orleans: Association for Computational Linguistics.

Maxwell E. McCombs, D. L. (1972). The Agenda-Setting Function of Mass Media. *The Public Opinion Quarterly*, 176-187.

Memmott, S. (2023, 9 6). *Research Guides: Misinformation, Disinformation, and Bias: News Media Bias*. Retrieved from Eastern Michigan University library: https://guides.emich.edu/misinformation/mediabias

Newman, N. (2023, 06 14). *Overview and key findings of the 2023 Digital News Report.* Retrieved from reuters institute of politics: https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2023/dnr-executive-summary

Nic Newman, R. F. (2023). *Reuters Institute Digital News Report 2023.* Oxford: Oxford University.

Ofcom. (2023). *News consumption in the UK: Supporting data.* Ofcom.

Oxford University Press. (2022). Do News Media Kill? How a Biased News Reality can Overshadow Real Societal Risks, The Case of Aviation and Road Traffic Accidents. *Social Forces*, 506–530.

Pascal Jürgens, B. S. (2022). Mapping Exposure Diversity: The Divergent Effects of Algorithmic Curation on News Consumption. *Journal of Communication*, 322-344.

Pew Research Center. (2023, 11 15). *Social Media and News Fact Sheet*. Retrieved from Pew Research Center: https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/

Psycology Today. (NA). *Bias | Psychology Today*. Retrieved from Psycology Today: Bias | Psychology Today

reuters institute. (n.d.).

Reuters Institute. (2023). *Digital News Report 2023*. Retrieved from Reuters Institute for the Study of Journalism: https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2023

Rosenstiel, B. K. (2010). *Blur: How to Know What's True in the Age of Information Overload.* New York City: Bloomsbury Publishing USA.

Stroud, N. J. (2011, 05 9). *Niche News: The Politics of News Choice*. Retrieved from Oxford Accademic: https://academic.oup.com/book/32769

Torbitt, C. (2024). How Daily Mail went from voice of Middle England to Tiktok sensation. *Press Gazette*.

Wihbey, J. (2015, 06 20). *Bias in news reporting is a significant issue that impact how audiences perceive and understand the news. News organisations and journalists strive to adhere to a code of ethics to ensure fair and reliable reporting, but biases can still appear in a varie*. Retrieved from The Journalist's Resource: https://journalistsresource.org/race-and-gender/racial-bias-reporting-research-trends/

wikipedia. (n.d.). *Bias*. Retrieved from Wikipedia: https://en.wikipedia.org/w/index.php?title=Bias&oldid=1204052935