**Research Proposal Presentation Transcript**

Hello, my name is Vasilisa Lukashevich. I am currently studying in the Computing Department at the University of Essex, pursuing an MSc in Artificial Intelligence Programme. As part of the Research Methods and Professional Practice module, which began in May 2023, I have prepared a Research Proposal Presentation on the topic of "Comparative Analysis of Political Bias Detection Results in the 'HonestyMeter' and 'Bias-o-meter' AI Tools." Likely, the work presented here could serve as the first step for my Capstone project, making feedback highly appreciated.

**Slide 1. Project Title**

Both of these tools offer open access and are freely available for use. On the first slide, you can view the logos of these bias-detection instruments. The relevant links and references will be listed at the end of the presentation.

**Slide 2. Contribution to the discipline**

In democratic countries, political knowledge is considered to be a cornerstone of the state (Open brackets, Carpini & Keeter, comma, 2022 close brackets). People have access to essential information about politics and express their support or disapproval for candidates during elections. That is why it is crucial to present politically related facts objectively.

In the last century, the traditional method of fact-checking was the most popular approach to evaluate texts in the media and realise if there was any propaganda or misinformation. Resources like PolitiFact emerged, where professional journalists assessed the relative accuracy of political statements using a six-point scale (True, Mostly True, Half True, Mostly False, False, Pants on Fire).

**Slide 3. Contribution to the discipline 2**

In this century we have witnessed a significant increase in the amount of digital information on the Internet, estimated at more than 90 zettabytes globally after 2020 (Open brackets, Jang et al., comma, 2018, close brackets). Politically related texts spread and multiply rapidly across various online platforms. Some of these texts may contain complete falsehoods and bias.

The development of artificial intelligence techniques has introduced new approaches, however there is limited research comparing the results of two open AI tools on the same political news texts. How reliable could these instruments be?

We assume that it would be reasonable to compare the scores given to the texts by humans (for example, a group of professional journalists using the mean score given by them) and AI tools. Probably, this aspect could be considered as the next step in research.

Understanding the effectiveness and limitations of such bias-check programs could be valuable for journalists and social media users who wish to utilise new tools in their work or everyday life.

The findings could improve the understanding of political bias detection and develop more effective AI tools in the future.

**Slide 4. Research Question**

According to the Oxford dictionary, bias is a prejudice in favour of or against somebody or something. When we refer to political bias, we mean a wide range of manipulations, from propaganda and political misinformation to one-sided judgments about a subject without considering alternative views. The tools in question utilise deep learning-based algorithms to analyse texts and mark the degree of bias in the content.

In this research, we want to conduct a practical comparison of the "bias-sensitivity" of two AI-driven open bias checker programs for news articles. The main question is:

*In what scales do the results of two instruments vary from each other on the same texts (open brackets, and why, close brackets)?*

We will further explain why we placed "*why*" in brackets.

**Slide 5. Aims**

We will identify several topics of news articles where the results of the observed tools are either similar or different.

We aim to analyse and interpret the differences in the results, exploring the most and least bias-sensitive topics in modern politics.

**Slide 6. Objectives**

The objective of our research is to compare two artificial intelligence tools, evaluating their effectiveness, similarities, and differences in detecting bias in political news texts.

We are going to briefly discuss the reasons for the mismatched results of the two AI tools. For instance, we may conclude that the data sources used to train the machine learning models were skewed, or there was inappropriate documentation and motivation of the dataset, or there is a weak connection to the current political context. (Open brackets, Paullada et al., comma, 2021, close brackets).

We will investigate basic reasons why these models may demonstrate *inherent* political bias.

To enhance understanding we will create visualisations that illustrate the differences and delve into the gap between the results obtained from the two projects.

**Slide 7. Key literature related to the project**

We propose four main literature sections, excluding the news texts corpus themselves, to support our experiment and its results.

**Section 1. The literature on political biases in news texts,** including machine learning algorithms practical implementations. It helps us better understand the special language and patterns for political bias detection. Please, look at the illustration on the left. For example, in US media, the phrase "Big Oil" may signal a high probability of criticising the establishment's position, while the expression "oil producers" indicates a pro-establishment bias (Open brackets, D’Alonzo & Tegmark, comma, 2022, close brackets).

On the other hand, certain phrases, such as "probably," "just”, "suggest" introduce uncertainty into the text, which has an impact on the results. In the illustration on the right Rashkin et al. (2017) provided an example from PolitiFact, where the use of "misleading phrasing was one reason for the "in-between ratings." Instead of definitive labels like "False" or "True," these mitigating words resulted such as "Mostly False" and "Mostly True."

**Slide 8. Section 2. The literature on complex political bias detection algorithms like deep learning, transformers and hybrid models.**

The authors of HonestyMeter highlighted that their intelligent agent can identify over 100 known manipulation techniques in texts. Bias-o-meter claims to "uncover hidden biases in the news". However, both of these programs employ not only a single Machine Learning Natural Language Processing (NLP) technique but rather a complex system of neural networks. The literature of this section helps us to understand that this approach resembles a classic black box, where it is difficult to explain why the output appears in a particular way.

However, there have been attempts among scholars to address this problem, such as implementing Knowledge Graphs to enhance the transparency and interpretability of deep learning systems (Open brackets, Gaur et al., comma, 2021, close brackets).

Bias-o-meter also emphasises that the program utilises transformers, which are large-scale pre-trained models like BERT from Google or GPT from OpenAI, for contextual comprehension and language generation. However, some scholars have noted that the original BERT model is considerably undertrained and have suggested alternative variants such as RoBERTa or FakeBERT. (Open brackets, Liu et al., comma, 2019, semicolon, Kaliyar et al., comma, 2021, close brackets). Once again, the utilisation of GPT algorithms leads to the model becoming an unexplainable black box, making it challenging to provide a precise explanation for the obtained scores.

**Slide 9. Section 3. The literature on the construction of bias-checker projects.**

We must confess that we do not have access to the complete code of both tools under consideration. Furthermore, there is currently no academic paper available that provides a comprehensive description of the HonestyMeter or Bias-o-Meter. This limitation significantly reflects on our ability to conduct a thorough evaluation and obtain a research-based answer to the question "why" we discussed earlier.

However, we suppose that while having limited access to the code may restrict a more in-depth analysis, a comparative evaluation based on user experience and observable characteristics can still provide valuable insights into the performance and usability of the bias detection tools.

Also, we can explore other similar projects in an effort to identify general vulnerabilities that may lead to varying and highly subjective output results. One such project is the Dbias, which assists in detecting and mitigating biases in NLP tasks. Dbias employs diverse machine learning techniques to enhance biased text until it produces a version that is either unbiased or at least more neutral. For our experiment we pay attention only for bias-checking algorithms from this system.

Researchers have emphasised the importance of continuously fine-tuning the bias detection module using different models and embeddings. Nevertheless, the Dbias package achieves a maximum accuracy of 77%. (Open brackets, Raza, comma, 2022, close brackets).

**Slide 10. Section 4. The literature on the quantitative research.**

To compare the outcomes of two tools, we rely on literature on statistics, for instance, “Basic Business Statistics” book by Berenson et al. (2015), especially its Chapter 10 on providing Two-Sample Tests.

However, we bear in mind that as researchers we may mistakenly perceive ourselves as playing a passive role in representing an existing reality via quantitative research. (Open brackets, Zyphur & Pierides, comma, 2017, close brackets).

To ensure credibility, the involvement of two researchers is important for quality rating of the studies to assess result validity. (Open brackets, Claydon, comma, 2015, close brackets) As this is an independent work, we once again see our limitations.

**Slide 11. Ethical considerations and risk assessment**

According to UNESCO, we need to be aware that AI algorithms are not neutral. AI tools can produce inaccurate results, lead to discriminatory outcomes, and contain embedded biases.

In our case the results may reveal inappropriate conclusions for both bias-detecting projects, highlighting their weaknesses and potentially impacting their reputation.

On the flip side, we acknowledge the risk that the results generated by AI tools may not be suitable for authors in the news media.

We propose that the responsibility for the output lies with the creators of the AI tools, as they have the autonomy to choose the datasets for training, feature selection, and testing the system (Open brackets, Dignum, comma, 2020, close brackets). On our part, our responsibility lies in conducting a thorough analysis of the scores difference.

**Slide 12. Ethical considerations and risk assessment 2**

Another risk involves including political discussions, which can be highly contentious when conducting research. Two critical perspectives on political ethics can be identified: the first argues that ethics has no place in politics, while the second suggests that political ethics excessively focuses on specific policies and policymakers (Open brackets, Thompson, comma, 2018, close brackets). We are aware that this research may predominantly align with a political realist or another perspective, and we take responsibility for that risk.

As we mentioned in the previous slides, this is personal academic research, it is essential to address potential biases introduced by the researcher, such as the subjective selection of texts for the comparison.

Therefore, careful consideration should be given to minimise any potential biases in order to maintain the integrity of the research.

**Slide 13. Methodology**

We compare the results from the users' perspectives.

By feeding the HonestyMeter and Bias-o-meter AI tools with approximately 50-100 politics-related texts written in English from different media sources in Great Britain, the U.S., Russia, and China. Then we record the results obtained from both tools in a table. Here is an example of analysing the same text, and we can see that the scores results look quite different.

Please, pay attention that there is a discrepancy in the scoring formats used by two tools. Bias-o-meter provides scores on a scale of 0 to 10, where 0 represents unbiased texts and 10 signifies highly biased articles. The other tool shows a 0% to 100% scale, with 0% indicating biased texts and 100% representing unbiased content.

**Slide 14. Standardising the results**

To ensure consistency, we need to transform the results into a unified format using the following formulas:

*HonestyMeter = 100 - Score Result Bias-o-meter = Score Result \* 10*

There are several approaches to comparing the results. For instance, we can assess the extent to which the results vary across news articles from different media sources, countries, topics, or individuals.

**Slide 15. Example of practical comparison**

A small example. We have a ready-made table with ten processed news texts and already calculated results according to our comparison methodology. You can see the two columns with the results. Additionally, we have saved detailed descriptions of the verdict from both applications for each of the news articles.

**Slide 16. Example of practical analysis**

Despite the limited number of texts, we can already draw some preliminary conclusions about the results of two AI tools. The mean of the results for Bias-o-meter is significantly higher than the mean of HonestyMeter, 51 versus 31. This probably means that the algorithms in Bias-o-meter are configured to be more sensitive, although this is not the case for all texts. In one text in the population, both instruments provided identical results.

**Slide 17. The initial pitfalls in practical work**

To present reasonable criticism and to briefly outline some new constraints that have emerged from the initial practical experience in analysis, let us consider a few. For example, sometimes both systems yield slightly different results for the same texts.

**Slide 18. The initial pitfalls in practical work 2**

One of the AI tools mitigates the overall outcome. For instance, according to the chart, in one of the texts, a 100% bias towards one side was detected, yet the final result is 50 out of 100.

**Slide 19. The initial pitfalls in practical work 3**

And finally, we encountered some technical issues with both AI tools as they struggled to process certain texts due to various reasons. This complicates the investigation of these systems.

**Slide 20. Timeline of proposed activities**

Allow me to conclude by briefly presenting a plan that outlines the time required to conduct my research project. We need approximately more than 300 hours of work. Thank you!

1. Compile the dataset of several dozen texts written in different countries, sources, and angles (12-24 hours).
2. Process them through two bias-detecting tools (24+ hours).
3. Analyse/Interpret the results (72+ hours).
4. Generate graphics (48-72 hours).
5. Explain why the disparities between results may arise (24-48 hours).
6. Write a document (72+ hours).

**Slide 21. References**

Anderson, M. (2022) An AI-Driven Bias Checker for News Articles, Available in Python. Available from: https://www.unite.ai/an-ai-driven-bias-checker-for-news-articles-available-in-python/) [Accessed 28 Jun 2023]

Berenson, L., Levine, D. & Szabat, K. (2015) *Basic Business Statistics: Concepts and Applications*. 13th Ed. Pearson

Bias-o-meter. Available from: https://bias-o-meter.optic.xyz/ [Accessed 3 Jul 2023]

Claydon, L.S. (2015) Rigour in quantitative research. *Nursing Standard (2014+), 29(*47), p.43.

D’Alonzo, S. & Tegmark, M. (2022) Machine-learning media bias. *Plos one, 17*(8), p.e0271947.

Dignum, V. (2020) Responsibility and artificial intelligence. *The oxford handbook of ethics of AI*, *4698*, p.215.

HonestyMeter. Available from: https://www.honestymeter.com/ [Accessed 3 Jul 2023]

Gaur, M., Faldu, K. & Sheth, A. (2021) Semantics of the black-box: Can knowledge graphs help make deep learning systems more interpretable and explainable?. *IEEE Internet Computing, 25(*1), pp.51-59.

Hsu, T. & Lee Myers, S. (May 18, 2023) Another Side of the A.I. Boom: Detecting What A.I. Makes. *The New York Times*. Available from: https://www.nytimes.com/2023/05/18/technology/ai-chat-gpt-detection-tools.html [Accessed 28 Jun 2023]

Jang, W. J., Kim, J. Y., Lim, B. T., & Gim, G. Y. (2018). A study on data profiling based on the statistical analysis for big data quality diagnosis. *International Journal of Advanced Science and Technology, 117*, 77-88.

Kaliyar, R.K., Goswami, A. & Narang, P. (2021) FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia tools and applications, 80*(8), pp.11765-11788.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019) RoBERTa: A robustly optimised BERT pretraining approach. *arXiv preprint arXiv:1907.11692.*

MTS Staff Writer (Apr 24, 2023) HonestyMeter: AI-Powered Pioneer for Unbiased Media Integrity and Transparency. *Marketing Technology Insights*. Available from: https://martechseries.com/predictive-ai/ai-platforms-machine-learning/honestymeter-ai-powered-pioneer-for-unbiased-media-integrity-and-transparency/ [Accessed 28 Jun 2023]

Oxford English Dictionary. Available from: http://www.oed.com/ [Accessed 3 Jul 2023]

Paullada, A., Raji, I.D., Bender, E.M., Denton, E. & Hanna, A. (2021) Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns, 2*(11).

PolitiFact. Available from: https://www.politifact.com/ [Accessed 3 Jul 2023]

Rashkin, H., Choi, E., Jang, J.Y., Volkova, S. & Choi, Y. (2017) Truth of varying shades: Analysing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2931-2937).

Raza, S., Reji, D.J. & Ding, C. (2022) Dbias: detecting biases and ensuring fairness in news articles. *International Journal of Data Science and Analytics*, pp.1-21.

Thompson, D.F. (2018) Political ethics. International encyclopaedia of ethics.

UNESCO (21 Apr, 2023) Artificial Intelligence: examples of ethical dilemmas. Available from: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics/cases [Accessed 4 Jul 2023]

Zyphur, M.J. & Pierides, D.C. (2017) Is quantitative research ethical? Tools for ethically practicing, evaluating, and using quantitative research. *Journal of Business Ethics, 143*, pp.1-16.

**Slide 22. References for texts under experimentation**

Aliaj, O. (Jul 4, 2023) Trump Media merger partner reveals $18mn preliminary settlement with SEC. *Financial Times*. Available from: https://www.ft.com/content/84ff4069-623d-41e2-99d2-08f3c0724c8d [Accessed 4 Jul 2023]

Caldwell, L. A., Dawsey, J. & Sanchez, Y.W. (Jul 1, 2023) Trump pressured Arizona Gov. Doug Ducey to overturn 2020 election. *The Washington Post.* Available from: https://www.washingtonpost.com/nation/2023/07/01/trump-2020-election-arizona-governor-doug-ducey/ [Accessed 4 Jul 2023]

CGTN News (Jun 27, 2023) China ratifies WTO deal on fishery subsidies. *CGTN*. Available from: https://news.cgtn.com/news/2023-06-27/China-ratifies-WTO-deal-on-fishery-subsidies-1kYTGFW7gac/index.html [Accessed 4 Jul 2023]

He, L. (Jul 4, 2023) China just played a trump card in the chip war. Are more export curbs coming? *CNN*. Available from: https://www.cnn.com/2023/07/04/tech/china-export-controls-semiconductor-war-explainer-intl-hnk/index.html [Accessed 4 Jul 2023]

Huynh, A. (Jun 25, 2023) Trump Campaigns in Michigan, a Battleground That’s Tinted Blue. *The New York Times*. Available from: https://www.nytimes.com/2023/06/25/us/politics/trump-republicans-michigan.html [Accessed 4 Jul 2023]

Kilander, G. (Jul 4, 2023) Truth Social SPAC agrees to SEC settlement as Trump’s past comments on indictment resurface. *Independent*. Available from: https://www.independent.co.uk/news/world/americas/us-politics/trump-news-indictment-desantis-truth-social-latest-b2368968.html [Accessed 4 Jul 2023]

Maruf, R. (Jul 3, 2023) SPAC seeking merger with Trump’s media company agrees to settlement with SEC. *CNN*. Available from: https://www.cnn.com/2023/07/03/business/trump-spac-sec-settlement-in-principle/index.html [Accessed 4 Jul 2023]

Ndegwa, S. (Sep 10, 2020) Trump hates black people. Period. *CGTN*. Available from: https://news.cgtn.com/news/2020-09-10/Trump-hates-black-people-Period--TFwAGRHdKM/index.html [Accessed 4 Jul 2023]

RT (Jul 2, 2023) LGBT Republicans slam DeSantis ad as ‘homophobic’. *RT.com*. Available from: https://www.rt.com/news/579106-desantis-video-homophobic-republicans-lgbt/ [Accessed 4 Jul 2023]

Smith, D. (May 13, 2023) As Trump’s lies and scandals deepen, the GOP responds as usual – with silence. *The Guardian*. Available from: https://www.theguardian.com/us-news/2023/may/13/trump-lies-republicans-congress-mccarthy-mcconnell [Accessed 4 Jul 2023]