EPISTEMOLOGICAL BIAS AS A MEANS FOR THE AUTOMATED DETECTION OF INJUSTICES IN NEWS MEDIA

Anonymous authors
Paper under double-blind review

ABSTRACT

Injustice occurs when someone experiences unfair treatment or their rights are violated. In the context of news media, injustices represent a form of bias through which discriminatory narratives can arise and spread. The automated identification of injustice in text has received little attention, due in part to the fact that underlying stereotypes are rarely explicitly stated and that instances often occur unconsciously due to the pervasive nature of prejudice in society. Here, we leverage the combined use of a fine-tuned BERT-based bias detection model, two stereotype detection models, and a lexicon-based approach to show that epistemological biases (i.e., words, which through their use, presupposes, entails, asserts, hedges, or boosts text to erode or assert a person's capacity as a knower) can assist with the automatic detection of injustice in text.

1 INTRODUCTION

The most basic duty of the media is knowledge sharing. Yet, the tool necessary for wide-spread knowledge sharing is influence. With this influence, the media is able to shape how one will understand the intricacy of a storyline in a news story with little effort. This often results in the use of epistemological biases which involves propositions that are presupposed, entailed, asserted, hedged, or boosted in text Recasens et al. (2013) to erode or assert a person's capacity as a knower, leading to framing issues and injustice within the text. Particularly, some of these word choices lead to testimonial injustice and what we define as character injustice. Character assassination is "the deliberate destruction of an individual's reputation or credibility" Icks et al. (2019). This often leads to character injustice, which is an unjustified attack on a person's character that results in an unfair criticism or inaccurate representation of them. Testimonial injustice occurs when modified believability is assigned to the statement of a subject based on widely known stereotypes Fricker (2007). Both of these types of injustices can lead to affirming or perpetuating stereotypes concerning the subject. Though it has always been a harsh reality with various ringing consequences, in recent years we have publicly witnessed how the affirmation of stereotypes can lead to physical violence, prejudice, and negative self-image [Harrison & Esqueda (1999), Gover et al. (2020), Kuykendall (1989)]. These experiences are harmful and dangerous, explicitly for the victims but for all members of our society. We also consider framing injustice. According to Entman (2007) framing bias happens when the use of subjective, one-sided words that reveals the stance of an author occurs. Which means that an individual's choice from a set of options is influenced more by *how the information is worded*, rather than by the information itself. We recognise that news content can be positively or negatively framed to influence the narratives. In this paper we aim to show how the negative cases when it affects particular subjects or individuals, can lead to framing injustice occurring towards those subjects.

In this work, we seek to detect character, testimonial, and framing injustices with a framework which includes a fine-tuned BERT model based on work from Pryzant et al. (2020b) to automatically tag words associated with epistemological bias from an input text, use the Kwon & Gopalan (2021) CO-STAR model and Sap et al. (2020) Social-Bias Frames to find stereotypes and the concepts of those stereotypes associated with the input text, and show when the tagged words (associated with some epistemological bias) or less credibility of a person are correlated with a stereotype which

causes injustice. We will use news media as a use-case throughout this work. Thus, we present the following contributions:

1. We develop a framework which uses results of 3 models to detect character, testimonial, and framing injustices in News Media.
2. We produce a fine-tuned tagger model to automatically detect epistemological bias.
3. We develop a User-Interface for journalist and editors to submit text to and receive output and explanations surrounding the tagged word, referred to in this work as tagger-UI.
4. We produce evidence showing epistemological bias is correlated with injustices in news article headlines.

One goal of our work is to give journalists and editors a tool which will help them learn how to avoid character, testimonial, and framing injustices in their work. This will be accomplished by showing users which words they use that produce epistemological bias, showing them the potential stereotype associated with the tagged words and text, offering the user explainability with the help of the stereotype concepts as defined by Kwon & Gopalan (2021), offering the user resources to reference literature on the particular epistemological bias identified in their input text.

2 BACKGROUND

Many works have established it is difficult for the common person to identify a biased word in a sentence and establish the need for computational agents to take on this charge Recasens et al. (2013). The difficulty arises due to us holding our own biases as facts and lack of education on sentence construction. This difficultly also exists since the framing of a sentence and phrases (sets of words) reveal injustice rather than a single word choice. However, there are specific word choices which are epistemologically biased and can lead to injustices occurring. We will focus on these word choices throughout this work.

2.1 WORKS ON EPISTEMOLOGICAL BIAS DETECTION

Great efforts have been put towards identifying potential words in text materials that could encourage epistemological bias [Recasens et al. (2013), Hube & Fetahu (2019), Pryzant et al. (2020b)]. Following on from the work of Pryzant et al. (2020b) we fine-tuned their tagger model to automatically tag words associated with epistemological bias from an input text. Identifying words which cause epistemological bias is a step towards awareness of social harms. This begs the question, what do we do with our new found knowledge and awareness? What kinds of implications does our use of these words impact society? What communities are affected by these word choices? These are the questions we explore in this work.

Authors of Kwon & Gopalan (2021) have trained a model to detect widely-known stereotypes and the concept of those stereotypes in text materials. We leverage the results of their CO-STAR model and the Social Bias Frames model Sap et al. (2020) to offer some explainability of the word choices by the model. Associating a particular text with a stereotype and the concept of that stereotype is a critical step towards awareness of social harms that might cause character injustice to a particular individual or group. Lack of identifying the words in a sentence which imply and promote these stereotypes leaves us with the undirected burden and question of: how can we address these harms? This will be further discussed in the methods section of the paper.

2.2 WORKS ON TESTIMONIAL INJUSTICES

In Beach et al. (2021) the authors identify words in text that cause testimonial injustice in medical records of Black patients. Detecting such testimonial injustice is helpful in seeing the unjust realities of our society. Beach et. al conclude the testimonial injustice that persists in these medical records has a high potential of causing disparity in quality of healthy care for Black patients, which correlates with findings that Black patents receive worse quality health care Odonkor et al. (2021). Identifying testimonial injustice in text materials is vital to creating an environment of accountability. For accountability to be applied we must include education, which is often unexplored or left up to the

user. Many users do not know where to find resources for such things, thus we provide them in our framework.

2.3 WORKS ON CHARACTER ASSASSINATION

Character assassination rears itself in character injustice. In Marouf et al. (2019), they detect character assassination from the comments of trolls from social media posts. Though they acknowledge the harms of character assassination, there is very little incentive to offer explainability in social media contexts since the goal of trolls is to cause frustration. We seek to address our users and educate them since in news media there is a benefit in preventing this pattern of behavior.

2.4 OTHER WORKS ON DETECTING BIAS

In Raza et al. (2022) authors developed a pipeline which takes in news articles, detects and masks words that are biased, and suggests words with more neutral text. Whilst the pipeline and library designed by the authors is very good and useful, they however don't consider the linguistic and epistemological bias features as discussed in Recasens et al. (2013) and used by Pryzant et al. (2020b). Unlike our work discussed in this paper, their pipeline produces no way of distinguishing and highlighting epistemological bias types of any identified potential biased word in the sentence. Also, their framework does not attempt to relate potential stereotypes and stereotype concepts that might be associated with causing injustice to a person or group.

3 METHODOLOGY

Hamborg et al. (2019) concluded that various forms of media bias is already analysed in the Social Sciences field and can be implemented in an automated fashion by primarily using Natural Language Processing (NLP) techniques. We leverage research methods using NLP and deep learning whilst also using analysis concepts by researchers from social sciences, who have studied media bias for decades.

We propose a technical framework which includes NLP models for detecting potential epistemological biased words and potential stereotypes along with their concepts, which we semantically rank, in a given news article headline. We do this with an aim of showing how the automated detection of epistemological bias can help us find injustices in news text. This framework is made up of but not limited to the attributes discussed here.

3.1 TAGGER MODEL

We leverage the detection component of the modular model as discussed in Pryzant et al. (2020b) and refer to this as the *tagger model* for the rest of this paper. The tagger model takes as input a given sentence (e.g. a headline text) and predicts the probability of each word in that sentence been biased. Then it returns the word with the highest probability as the tagged biased word. The detection model itself is a BERT-based neural sequence tagging model that has been fine-tuned to include the expert linguistic and epistemological bias features from Recasens et al. (2013). The model leverages the semantic meaning of each word in the given sentence via the BERT Kenton & Toutanova (2019) contextualised word vector. We extend the detection model further by including not only the expert linguistic and bias features but also a feature on whether or not the sentence contained too much information (TMI). This TMI information looks at instances where news article headlines include information which does not add to understanding the news article, but seeks to cloud the judgement of the user. We hypothesised here that - the TMI information can be considered from a linguistic standpoint to be a sentence instance which has more than 2 describing words in it and in English language there are mainly two types of describing words, i.e. adjectives and adverbs.

We trained the model with a learning rate of 3e-5 and the extra features concatenated on top of the BERT model. With the pre-enrich option, we are able to pass the features through a linear layer before they get added in. We also made use of a debiasing weight option of 1.3 as opposed to the default of 1.0 used in Pryzant et al. (2020b).

3.2 CO-STAR MODEL AND SOCIAL BIAS FRAMES (SBF)

The CO-STAR (COnceptualisation of Stereotypes for Analysis and Reasoning) framework allows input from the user and generates outputs of stereotypes and stereotype concepts. The Social Bias Frames model generates and classifies stereotypes associated with an inputted text. These outputs are quite simple to understand, but have very complex history.

Since testimonial injustice occurs because of widely known stereotypes, the outputs of these models will help us inform our users of any potential stereotypes that are exacerbated by their inputted text. The potential stereotype can also be used to determine if a person's character has been unjustly targeted. This is because stereotypes are not based in the actual truth of the particular person they are attached to and stigmatises the individual. There presence amidst character assassination is evidence of character injustice. Stereotype concepts will offer explainability as to why a certain word was tagged as being epistemologically biased.

We submit news article headlines to the models and receive an output of potential stereotypes and stereotype concepts the sentence is related to. We rank these outputs based on their semantic similarity to the headlines. The stereotype and stereotype concept most correlated with the sentence is output to the user as the potential stereotype and concept which casts doubt on the subject of the inputted text.

3.3 LEXICON LOOKUP

Once the tagger model has returned for us the top tagged word, we then proceed to automatically look up and semantically search the tagged word in the epistemological lexicons from social sciences - to discover the epistemological bias types it is associated with. These lexicons are from the collection of datasets we discuss in the dataset section 3.5. When a tagged word is not found in the lexicons, we lemmatize said word (considering its context) to find its base word and find the stem word (removing the prefixes and suffixes). We then search for the lemmatized and stemmed words in the lexicons. If the confidence of the tagged word being epistemologically biased is greater than 50%, and the word nor its lemma nor its stem is found, we consider the tagged word to be a regular word not associated with any of the lexicons.

3.4 INTERACTIVE INTERFACE FOR LEARNING ABOUT BIAS

We propose and develop the use of an interactive interface for journalist and editors of news media text content to learn about any potential bias types and associated stereotypes that might exist in their content, prior to them publishing it. The intention is not to only point out potential biased words, but to also provide them with the knowledge and resources to any associated epistemological bias types. Which will help facilitate their ability to be more educated about the potential bias types and their effects, injustices they cause, and suggestive ways they can mitigate them from occurring.

Once logged in to the interface, the user gets presented with a landing page. This landing tells the user what kind of information they can input and gives instructions on how to input the information. Once ready, the user can paste in their material and simply hit on an analyze button to begin the process.

During the analysis, we first leverage the nltk python library to remove any stop words (i.e. a, the, etc..). We then check the sentence length. If the sentence is less than 3 words, we determine there is not enough context to analyze the text for potential stereotypes. The assumption here is that - though a sentence can be complete with 1-2 words, it is not likely to give much context. Thus, the system looks for sentences with greater than 3 words to analyze. Here, we are met with our first output option (a). A message is displayed to the user stating they have not provided enough context to classify their content and they should enter a longer sentence. However, if we have enough context, the content is sent to the epistemological tagger model to find the word with the highest probability of bias. We now explore our second output option (b). If the tagged word's probability of being epistemologically biased is less than 10%, we display to the user there is no bias detected in the submitted content. When the probability of the word being bias is greater than 10%, we automatically look for the tagged word in the epistemological bias lexicons as discussed in our methods: lexicon lookup section 3.3. If the word is found in any of the lexicons, we capture which lexicons it is associated. Whilst, if the word is not found in at least one of our lexicons, the tagged

term is referred to as a *regular* biased word. We display the types of associated epistemological bias detected to the user and provide links to resources which explains more on the specific types of epistemological biases detected - this is our third output option (c). Also, if the probability of the word being biased is greater than 50%, the content is sent to the CO-STAR and Social Bias Frame model, to be associated with a potential stereotype and concept of the stereotype. By semantically ranking the potential stereotypes and stereotype concepts to the sentence, we are able to display the most likely potential stereotype and stereotype concept associated with it to the user - our final output option (d). See examples of these outputs in table 6 and table 5 of the appendix section.

Sentence No.	Headline	Subject
1	Meghan Markle spent a staggering £38,000 on her clothes for a charity trip	Meghan
2	Kate Middletons £100,000 Astonishing value of the dress that won a Prince's heart (and has hung in a wardrobe for eight years)	Kate
3	Meghan Markles beloved avocado linked to human rights abuse and drought, millennial shame	Meghan
4	Kates morning sickness cure? Prince William gifted with an avocado for pregnant Duchess	Kate
5	Kate and William 'packed up the kids' in search of 'privacy' at new Windsor estate	Kate
6	Why Prince Harry and Meghan Markle's Potential Plan to Protect Their Family is Incredibly Unfair to Archie	Meghan
7	Prince Harry and Duchess Meghan Think Moving to Canada Will Give Archie a Normal Upbringing	Meghan
8	Not long to go! Pregnant Kate tenderly cradles her baby bump while wrapping up her royal duties ahead of maternity leave - and William confirms she's due 'any minute now'	Kate
9	Why can't Meghan Markle keep her hands off her bump? Experts tackle the question that has got the nation talking: Is it pride, vanity, acting - or a new age bonding technique?	Meghan
10	Kate Middleton Wore a Bardot Dress to the 'Top Gun' Premiere	Kate
11	Kate Middleton's homegrown bouquet of lily of the valley follows royal code	Kate
12	Royal wedding: How Meghan Markles flowers may have put Princess Charlottes life at risk	Meghan
13	Duchess Kate reveals her favourite photo of son Prince Louis	Kate
14	How Meghan and Harry Ripped Up Royal Tradition on Birthday Photos of Archie	Meghan
15	Kate Middleton Debuts New Sapphire Earrings That Belonged to Princess Diana and Debunks a Rumor!	Kate
16	Meghan Markle Just Casually Rewore Her $16,500 Royal Wedding Earrings in NYC	Meghan
17	Kate and Wills Inc: Duke and Duchess secretly set up companies to protect their brand - just like the Beckhams	Kate
18	A right royal cash in! How Prince Harry and Meghan Markle trademarked over 100 items from hoodies to socks SIX MONTHS before split with monarchy - with new empire worth up to £400m	Meghan
19	Kate Middleton deviated from royal 'cleavage protocol' with daring neckline	Kate
20	Meghan Markle Just Wore a Plunging Gown With a Thigh-High Slit on the Red Carpet	Meghan

Table 1: Example Headline Sentences Used

3.5 DATASET SOURCE USED

We leverage and used the bias data corpus[1] as created by Pryzant et al. (2020b). It contains a Wikipedia neutrality corpus showing Wikipedia articles that were revised for neutrality by editors on Wikipedia, and it also contains some real world news samples. Each example data point contained linguistic features of the sentence as discussed in Recasens et al. (2013). We extended this dataset by including the extra feature of whether the line contains too much information (TMI) or not as discussed in section 3.1.

Lexicons used in the lexicon lookup were from social science research and collated in Pryzant et al. (2020a). We compiled these lexicons into one large dictionary and included some metadata about the lexicon i.e. source, creators, resources about the epistemological bias type, etc.

[1] http://nlp.stanford.edu/projects/bias/bias_data.zip

When it comes to Meghan Markle and Kate Middleton, the media depicts several aspects of their lives which they share differently and often can be reflective of framing or character injustices[2]. We use headlines (as seen in Table 1) from such depictions in our comparative test to further illustrate the detection and harms of these injustices. The results of this comparative test are further discussed in our results section 4.

4 RESULTS AND DISCUSSIONS

4.1 COMPARATIVE TEST

We performed a comparative test to illustrate how our framework shows injustices e.g. character and framing injustice. Particularly, we capture the type of epistemological bias attached to the tagged word in each entry, observe if a relevant stereotype is associated with the inputs, and observe the depiction of different subjects. In this study, we look at Meghan Markle and Kate Middleton as our subjects of interest. Even before Meghan Markle had resigned from her royal duties, the media has sought to minimize her experience and diminish her comments and character. We often see instances of Meghan as the subject of an article and the media using sarcasm and criticism towards her. Such acts have led readers to having tainted images of Meghan and even not believing her statements or actions, thus testimonial injustices. We also see this leading to her character been questioned, thus character injustice. Yet, for the same and similar topics of concern, media members speak charmingly about Kate, thus leading to framing injustices shown. It is important to note here, that Kate Middleton and Meghan Markle shared similar positions, interests, and abilities. With this, we see them as good subjects to observe in our comparative test.

No	Subject	Tagged Bias Word	Potential Bias Type	Sentence Sentiment
1	Meghan	['staggering']	['subjectives']	negative
2	Kate	['astonishing']	['positive', 'subjectives']	neutral
3	Meghan	['beloved']	['positive', 'subjectives']	negative
4	Kate	['gifted']	['positive', 'subjectives']	neutral
5	Kate	['packed']	['regular']	neutral
6	Meghan	['incredibly']	['positive', 'subjectives']	negative
7	Meghan	['normal']	['subjectives']	neutral
8	Kate	['confirms']	['entailments', 'report']	negative
9	Meghan	['vanity']	['negative', 'subjectives']	negative
10	Kate	['top']	['entailments', 'positive', 'subjectives']	neutral
11	Kate	['homegr']	['regular']	neutral
12	Meghan	['royal']	['regular']	neutral
13	Kate	['favourite']	['regular']	neutral
14	Meghan	['ripped']	['negative']	neutral
15	Kate	['bunks']	['negative', 'subjectives']	negative
16	Meghan	['casually']	['regular']	neutral
17	Kate	['just']	['subjectives']	negative
18	Meghan	['markle']	['regular']	neutral
19	Kate	['daring']	['entailments', 'implicatives', 'positive','subjectives']	positive
20	Meghan	['just']	['subjectives']	neutral

Table 2: A subset of output result. (It shows the words tagged as potentially biased, the lexicon that type of bias is associated with, and the overall sentence sentiment for each headline sentence depicted in table 1). For a full list of columns, please refer to the appendix A

In Figure 1, we plotted results of an experiment where we ran 10 article headlines of Meghan Markle and 10 articles of Kate Middleton through our framework. Each article discusses one topic and how the subject relates to the topic. You can see the headlines for each subject in Table 1. The complete result output and table column descriptions are given in Appendix A. In this plot, we capture the subject of the headline (Meghan/Kate), sentiment of the entire headline sentence, and the associated epistemological bias types for each headline as shown in table 2. The light blue circles represent the overall sentiment of each headline, we will refer to them as the sentiment headline level. We use the 3 categories of positive, neutral, and negative sentiments. Within each sentiment headline level, we capture the subject of the headline, we will call this the subject level. Note that in Figure1,

[2]https://www.buzzfeednews.com/article/ellievhall/meghan-markle-kate-middleton-double-standards-royal

Kate takes up the entire positive sentiment space, there were no examples that were detected to have positive sentiments for Meghan. Within each subject level, we capture the epistemological bias types associated with the tagged words in the headlines for each subject; we will refer to this as the bias-type level. This illustration shows us how much space is filled by each subject.

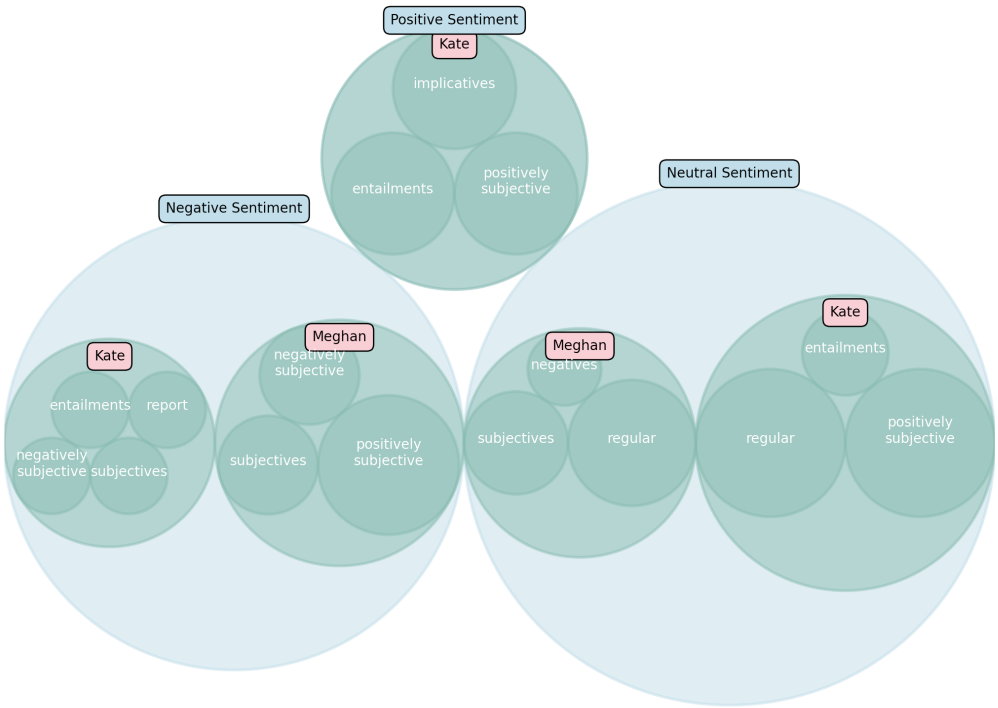

Figure 1: Epistemological Bias Types - Comparative Test

From this plot, we can see the headline sentiment level, the negative sentiment circle is mostly filled with articles about Meghan and the most common bias type is positively subjective. The intuition here is that many articles about Meghan contain sarcasm, though this is not the only indicator. There was one article about Kate which had negative sentiment and each of the bias types in her circle were identified in the headline, this is why they are all equal in size. In looking at the headline sentiment level, the positive sentiment circle is completely filled with articles about Kate and the most common bias type is implicatives. The intuition here is that, in articles which have common topics between Meghan and Kate, only Kate is talked about in a suggestively positive way. In looking at the headline sentiment level, the neutral sentiment circle is mostly filled with articles about Kate and the most common bias type is positively subjective and regular. Recall, regular is the epistemological bias type assigned to detected words which do not appear, nor similar words appear in the lexicon lookup. We can also observe, the articles which appear for Meghan with neutral sentiment are negatively subjective in nature, informing us they likely are not sarcastic but have negative suggestions. For Kate, she is talked about in a suggestively positive way, even in the neutral sentiments. It is clear from this experiment, when there is a topic that Meghan and Kate share - the media uses words for Meghan which are sarcastic and cause framing and character injustices to her as a subject. Both of which can contribute to someone taking Meghan less seriously, thus causing testimonial injustice.

We also captured the stereotypes which were associated with each headline. As discussed in section 3.2, relevant stereotypes and stereotype concepts can be generated, and we do so when the probability of the potential epistemological biased word obtained from the tagger model is above 50%. Looking at the potential stereotypes generated with a semantic similarity distance greater than 0.3 to the headline, as shown in table 3 (see appendix table 6 for the full table), we observe that 5 out of 7 of such entries related to Meghan as the subject and only 2 related to Kate. More interestingly, we observed that one of the potential stereotypes relating to Meghan (*personal spending habits*) aimed

No.	Subject	Stereotype(S)	Distance(S)	Stereotype Concept(SC)	Distance(CS)
1	Meghan	personal spending habits	0.3457	women should spend money on clothes	0.4914
2	Kate	women should be dressed like brides	0.3259	women are property	0.2278
3	Meghan	feminism	0.3663	sexism	0.2523
6	Meghan	family law	0.3244	arab folks are not protected	0.2951
10	Kate	women should wear dresses	0.3131	women should wear dresses	0.3131
18	Meghan	royalty housekeeping are gold diggers	0.3741	royalty house queens are gold diggers	0.3657
20	Meghan	women should dress modestly	0.3840	women should wear dresses that cover themselves	0.4047

Table 3: Table showing the potential top ranked stereotype and stereotype concept. We passed each sentence through the costar and social bias frames model. After which the potential stereotype and stereotype concepts where semantically ranked to the sentence itself

at a personal attribute. Which can be seen as an indication of a potential character assassination, thus character injustice.

4.2 Potential issues elevated due to the proposal and means to resolve them

A critical area for bias in systems and models' design often stem from a given human's intrinsic biases. They are usually a reflection of ourselves. One possible solution to this problem is to ensure that a diverse group of individuals are involved from the inception of the solutions design to the testing phase of such technical solutions—this is one reason why we gathered a diverse group of authors to be involved in the discussions presented in this paper.

A second area for bias surrounds the definition of terms and assumptions biases. A way we attempt to resolve this is by considering the existing consensus definitions of epistemological bias from the social sciences, where media bias has been studied for decades. Also, for our comparative test, we set out to have a baseline definition for injustice towards a subject as follows:

> *injustice in the media towards a subject can manifest as news that attempts to portray that subject in a negative light more than it portrays in a negative light another subject of similar standing due to subtle differences such as race, age, religion, etc.*

A third potential area for bias can occur in the training and validation datasets, as well as in the models used for implementing the proposed solution. For example, the sourced datasets do not contain enough semantic information that is very reflective of the injustices considered. One possible way to combat this is to investigate ways to build more robust training datasets or leverage models that do not require lots of training examples to generalise properly.

Broader problems may arise in any situation where technology is naively applied to solve a societal issue. As envisaged, our framework should be applied as a means to help people working in the media improve their output with respect to bias and injustices. However, as warned by Goodhart's law Manheim & Garrabrant (2018), if the measures and metrics suggested here become targets, they will cease to be useful. This is challenging to prevent and will require careful considerations before deployment of the solution in the wild.

5 Conclusions

We propose a framework which uses an epistemological tagging model, a stereotype detection model and social bias frame model, combined with semantic searching and lexicon lookup of epistemological biased word to detect character, testimonial, and framing injustice. These forms of injustice are often subtle and hard for them to detect or be aware of. We also provide empirical evidence for using this framework in media settings to detect such injustices. Further, we develop an interactive interface which can aide editors and journalists of news text content by automatically detecting and explaining potential bias types and injustices that might be present in their content. We anticipate this interactive interface will encourage them to take necessary, preventative steps in avoiding unjust acts before their content is realised.

6 FUTURE WORK

Below, we list the potential future extensions of our proposed solution.

- More Comparative Test data:- Our goal with the comparative test was to show how the tagger model, CO-STAR, SBF models and epistemological lexicons can be used together to highlight and educate editors and creators of news content on potential epistemological biases and injustices. However, we recognise that a limitation of the comparative test in this paper is in using only 20 headlines, and we hope to address this in the future by using more headlines over a time period for both subjects used in the comparative analysis.
- Learning Better Patterns for Testimonial injustice:- Identifying a single word which potentially cause epistemological bias in a given sentence is only the start. Bias is usually in the framing of an entire sentence and therefore, ability to tag phrases will further this work into more fully capturing biases. We do believe the principle of our framework to offer explainability will still be needed in this situation. Therefore, we hope to incorporate this capability in future works.
- Veridicality Assessment:- There are instances were words associated with causing epistemological bias can be factual in the context of a news article headline. We intend to incorporate veridicality assessment in our framework to assess if a statement is actually factual, in which case it should not be associated causing injustice or promoting bias. This will allow for a more accurate analysis of the entered text to the system.

7 REPRODUCIBILITY STATEMENT

We discuss how we leverage the models our framework and the outputs of the interactive interface in section 3 of this work. The datasets used in our work are discussed in section 3.5. Each dataset is publicly available, except one which is proprietary and used to tune the tagger model. However, the tagger model can be fine-tuned to a dataset which is relevant, ethically gathered, and appropriately constructed.

REFERENCES

Mary Catherine Beach, Somnath Saha, Jenny Park, Janiece Taylor, Paul Drew, Eve Plank, Lisa A Cooper, and Brant Chee. Testimonial injustice: linguistic bias in the medical records of black patients and women. *Journal of general internal medicine*, 36(6):1708–1714, 2021.

Robert M Entman. Framing bias: Media in the distribution of power. *Journal of communication*, 57 (1):163–173, 2007.

Miranda Fricker. *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press, 2007.

Angela R Gover, Shannon B Harper, and Lynn Langton. Anti-asian hate crime during the covid-19 pandemic: Exploring the reproduction of inequality. *American journal of criminal justice*, 45(4): 647–667, 2020.

Felix Hamborg, Karsten Donnay, and Bela Gipp. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4): 391–415, 2019.

Lisa A Harrison and Cynthia Willis Esqueda. Myths and stereotypes of actors involved in domestic violence: Implications for domestic violence culpability attributions. *Aggression and Violent Behavior*, 4(2):129–138, 1999.

Christoph Hube and Besnik Fetahu. Neural based statement classification for biased language. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pp. 195–203. Association for Computing Machinery, 2019. doi: 10.1145/3289600.3291018.

Martijn Icks, Eric Shiraev, Jennifer Keohane, and Sergei A Samoilenko. Character assassination: Theoretical framework. In *Routledge handbook of character assassination and reputation management*, pp. 11–24. Routledge, 2019.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, pp. 4171–4186, 2019.

Crystal Kuykendall. Improving black student achievement by enhancing students' self-image. 1989.

Teyun Kwon and Anandha Gopalan. Co-star: Conceptualisation of stereotypes for analysis and reasoning. *arXiv preprint arXiv:2112.00819*, 2021.

David Manheim and Scott Garrabrant. Categorizing variants of goodhart's law. *arXiv preprint arXiv:1803.04585*, 2018.

Ahmed Al Marouf, Rasif Ajwad, and Adnan Ferdous Ashrafi. Looking behind the mask: A framework for detecting character assassination via troll comments on social media using psycholinguistic tools. In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pp. 1–5, 2019.

Charles A Odonkor, Rachel Esparza, Laura E Flores, Monica Verduzco-Gutierrez, Miguel X Escalon, Ryan Solinsky, and Julie K Silver. Disparities in health care for black patients in physical medicine and rehabilitation in the united states: a narrative review. *PM&R*, 13(2):180–203, 2021.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. Neutralizing biased text lexicons. `https://github.com/rpryzant/neutralizing-bias/tree/master/src`, 2020a. Accessed: 2022-06-20.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pp. 480–489, 2020b.

Shaina Raza, Deepak John Reji, and Chen Ding. Dbias: Detecting biases and ensuring fairness in news articles. *International Journal of Data Science and Analytics*, 2022.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1650–1659, 2013.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5477–5490. Association for Computational Linguistics, 2020.

A APPENDIX

Table 5 in this section shows the full output of tagging each headline sentence used in the comparative test. The `No.` column corresponds to the sentence number as shown in table 1, `Subject` refers to the person subject the headline relates to, `Taggerout Bias` represents the tagged word from the sentence that could potentially be biased, `Taggerout Prob` refers to the probability of the tagged word been biased, `Taggerout in lexicon` refers to the results of checking whether the corresponding tagged word is observed in the lexicon of associated epistemological biased words. This offers some explanability for the potentially biased word. A **True** value for this column indicates that the word is indeed in the lexicon, whilst a **False** value indicates otherwise. `Bias Type` indicates what kind of epistemological bias type is associated with the tagged word. A **regular** bias type word indicates that the tagged word was not found in the epistemological bias lexicon. The `Sentence Sentiment` columns represents the overall sentiment associated with the headline sentence.

In Table 6, the `Stereotype (S)` column refers to the closest potentially associated stereotype. After passing each headline through the COSTAR and Social Bias Frame models as discussed in

section 3.2, we semantically ranked the list of potential stereotype to the headline sentence, and then selected the top ranked potential stereotype and stereotype concept - as seen in the column `Stereotype Concept(SC)`. The `S Distance` and `SC distance` columns shows the semantic similarity distance of the stereotype and stereotype concepts to the headline sentence.

Sentence No.	Headline	Subject
1	Meghan Markle spent a staggering £38,000 on her clothes for a charity trip	Meghan
2	Kate Middletons £100,000 Astonishing value of the dress that won a Prince's heart (and has hung in a wardrobe for eight years)	Kate
3	Meghan Markles beloved avocado linked to human rights abuse and drought, millennial shame	Meghan
4	Kates morning sickness cure? Prince William gifted with an avocado for pregnant Duchess	Kate
5	Kate and William 'packed up the kids' in search of 'privacy' at new Windsor estate	Kate
6	Why Prince Harry and Meghan Markle's Potential Plan to Protect Their Family is Incredibly Unfair to Archie	Meghan
7	Prince Harry and Duchess Meghan Think Moving to Canada Will Give Archie a Normal Upbringing	Meghan
8	Not long to go! Pregnant Kate tenderly cradles her baby bump while wrapping up her royal duties ahead of maternity leave - and William confirms she's due 'any minute now'	Kate
9	Why can't Meghan Markle keep her hands off her bump? Experts tackle the question that has got the nation talking: Is it pride, vanity, acting - or a new age bonding technique?	Meghan
10	Kate Middleton Wore a Bardot Dress to the 'Top Gun' Premiere	Kate
11	Kate Middleton's homegrown bouquet of lily of the valley follows royal code	Kate
12	Royal wedding: How Meghan Markles flowers may have put Princess Charlottes life at risk	Meghan
13	Duchess Kate reveals her favourite photo of son Prince Louis	Kate
14	How Meghan and Harry Ripped Up Royal Tradition on Birthday Photos of Archie	Meghan
15	Kate Middleton Debuts New Sapphire Earrings That Belonged to Princess Diana and Debunks a Rumor!	Kate
16	Meghan Markle Just Casually Rewore Her $16,500 Royal Wedding Earrings in NYC	Meghan
17	Kate and Wills Inc: Duke and Duchess secretly set up companies to protect their brand - just like the Beckhams	Kate
18	A right royal cash in! How Prince Harry and Meghan Markle trademarked over 100 items from hoodies to socks SIX MONTHS before split with monarchy - with new empire worth up to £400m	Meghan
19	Kate Middleton deviated from royal 'cleavage protocol' with daring neckline	Kate
20	Meghan Markle Just Wore a Plunging Gown With a Thigh-High Slit on the Red Carpet	Meghan

Table 4: Comparative Headline Sentences Used

No.	Subject	Taggerout Bias	Taggerout Prob	Taggerout in Lexicon	Bias Type	Sentence Sentiment
1	Meghan	['staggering']	0.999498	True	['subjectives']	negative
2	Kate	['astonishing']	0.999342	True	['positive', 'subjectives']	neutral
3	Meghan	['beloved']	0.997946	True	['positive', 'subjectives']	negative
4	Kate	['gifted']	0.877285	True	['positive', 'subjectives']	neutral
5	Kate	['packed']	0.478948	False	['regular']	neutral
6	Meghan	['incredibly']	0.998493	True	['positive', 'subjectives']	negative
7	Meghan	['normal']	0.689049	True	['subjectives']	neutral
8	Kate	['confirms']	0.997673	True	['entailments', 'report']	negative
9	Meghan	['vanity']	0.599388	True	['negative', 'subjectives']	negative
10	Kate	['top']	0.933422	True	['entailments', 'positive', 'subjectives']	neutral
11	Kate	['homegr']	0.820782	False	['regular']	neutral
12	Meghan	['royal']	0.551708	False	['regular']	neutral
13	Kate	['favourite']	0.838293	False	['regular']	neutral
14	Meghan	['ripped']	0.890869	True	['negative']	neutral
15	Kate	['bunks']	0.758531	True	['negative', 'subjectives']	negative
16	Meghan	['casually']	0.967407	False	['regular']	neutral
17	Kate	['just']	0.645452	True	['subjectives']	negative
18	Meghan	['markle']	0.583964	False	['regular']	neutral
19	Kate	['daring']	0.856008	True	['entailments', 'implicatives', 'positive', 'subjectives']	positive
20	Meghan	['just']	0.565576	True	['subjectives']	neutral

Table 5: Table showing the output of passing each headline through the model that tags the potentially biased word in the given sentence.

No.	Stereotype (S)	Distance(S)	Stereotype Concept(SC)	Distance(SC)
1	personal spending habits	0.3457	women should spend money on clothes	0.4914
2	women should be dressed like brides	0.3259	women are property	0.2278
3	feminism	0.3663	sexism	0.2523
4	british women are marginalized for a joke	0.2083	pregnancy	0.3918
5	N/A	N/A	N/A	N/A
6	family law	0.3244	arab folks are not protected	0.2951
7	arab folks should live in a constant state of worry	0.1618	racial hierarchy	0.0999
8	gender hierarchy	0.1178	women are often subjected to terms like "babies" and "pregnant"	0.2632
9	women are sexual objects	0.2602	misogyny	0.2574
10	women should wear dresses	0.3131	women should wear dresses	0.3131
11	women are property of men	0.1226	women are property	0.1494
12	women are sex objects	0.1676	women are vulnerable	0.2495
13	sexual abuse	0.0746	saudi arabians are pedophiles	0.1226
14	sexual assault	0.2091	royalty house guests are paedophiles	0.4691
15	women are sex objects	0.1637	women are property	0.1504
16	black folks want to marry rich people	0.2266	racial marriage	0.0869
17	women are property	0.2026	women are secret business partners	0.3732
18	royalty housekeeping are gold diggers	0.3741	royalty house queens are gold diggers	0.3657
19	sexism	0.1927	gender hierarchy	0.2284
20	women should dress modestly	0.3840	women should wear dresses that cover themselves	0.4047

Table 6: Table showing the potential top ranked stereotype and stereotype concept. We passed each sentence through the costar and social bias frames model. After which the potential stereotype and stereotype concepts where semantically ranked to the sentence itself