An Approach for Targeted Analysis of Topic Portrayals in Public Service Media Corpora

Kristen M. Scott
KU Leuven
Leuven, Belgium
kristen.scott@kuleuven.be

Felix Mercer Moss
British Broadcasting Corporation
London, United Kingdom
felix.mercermoss@bbc.co.uk

Abstract

BBC News is beholden to public service values in providing it's output. It is also exploring the use of recommendation and personalization technologies for news services. There is also an interest in using computational technologies for the monitoring of adherence to public services. Despite these potential uses, there is a dearth of computational metrics for monitoring the vast text-based output of the BBC News service. Here we propose a method that could be utilized by the editorial team as a high level signal for issues in their output. The method is intended for monitoring the portrayal of specific topics in a corpus of news stories. It utilizes NLP tools of coreference resolution and dependency parsing to identify words related to the topic and BERT Language Model contextualized word embeddings to assess those words against a topic. We find that the method provides results consistent with human labeling when used on a benchmark dataset. We then perform a case study using a corpus of recent English language BBC News politics stories, where we test for associations in the text that relate to known gender-based stereotypes and do not find evidence of those stereotypes in the corpus. The approach presented may be an effective tool to monitor reporting to identify instances of bias or stereotyping. We conclude with a discussion of potential limitations of the approach and planned future work to validate and improve the proposed method.

Introduction

The British Broadcasting Corporation (BBC) is a public service media organization, which is beholden to public service values. The BBC Public Purposes are set out in the Royal Charter (2016), which is the constitutional basis of the BBC and is renewed every 11 years. (Note: The agreement with the government does not cover the BBC's Object, Mission, and Public Purpose). The Public Purposes are as follows:

- To provide impartial news and information to help people understand and engage with the world around them.
- To support learning for people of all ages.
- To show the most creative, highest quality and distinctive output and services.
- To reflect, represent and serve the diverse communities of all of the United Kingdom's nations and regions and, in doing so, support the creative economy across the United Kingdom.
- To reflect the United Kingdom, its culture and values to the world.

In the context of BBC News, news editors are responsible for curating what news content is presented to readers on the UK service site. The editorial guidelines are designed to ensure the meeting of the public purposes. Editors design the 'front page' seen online and they populate 'related articles' sidebars that are presented in conjunction with articles the readers select. For some international service sites a recommender system tool is used for identifying related articles, however, the editorial team has final approval over the implementation of any such automated applications (Boididou et al. 2021; Piscopo, Panteli, and Penna 2019).

There is significant interest in the use of technologies across the BBC for the support of meeting these purposes, as well as additional business goals; here we focus on English language BBC News which provides daily media coverage, in the form of text and video content. Technologies are already a key aspect of the BBC's services, for example, website and app-based interfaces are already the access point for BBC News content. There is a further interest in the possibilities of artificial intelligence (AI) based personalisation technologies, which are commonly used by non-public media companies. However, there is also caution about the use of technologies such as recommender systems, particularly in terms of their potential impact on provision of the public services values (Ada Lovelace Institute 2022). Possible concerns includes the possibility of varying recommendation quality for people, or groups of people as well as the potential creation of information silos for news consumers. In this way there is a risk of recommendations in news media no longer providing the mandated breadth of information and diverse content. There is also an interest in the potential of using AI-based technologies for monitoring adherence to public purposes, including for monitoring the impacts of any newly introduced personalization technologies.

We report here the work of developing an approach for calculating metrics for use by the editorial team for monitoring the portrayal of specific topics across BBC news services. We do this be identifying sentiment words within a

text corpora that are localized to the topic of choice. This has potential for use for monitoring purposes, to ensure that served content is reflecting diverse audiences and perspectives. We do this by utilizing contextualized word embeddings generated by a large language model (BERT) and exploiting the feature of these embeddings of similar texts to be closer in embedding space than dis-similar texts. We test the method using data labelled with positive and negative sentiment towards subjects of natural language text (news stories). We then apply the method to a corpus of BBC News text to assess for the existence of gender-based stereotypes within the text. We propose a flexible method to allow for users to expand measurement beyond sentiment, such as positive and negative. We allow users to be able to define the framework against which they want to measure the treatment of a specific topic, or concept (such as gender) in a corpora of text. This method is valuable for comparing multiple corpora, allowing for monitoring of changes in content over time, or of the content being served to different groups of people, to name some examples.

Our objectives in the following work are two-fold, firstly to demonstrate that our proposed method works for identifying the framing of a concept in the text and secondly, to utilize the method to ask some questions of a corpora of recent BBC news stories.

Related Work

There is extensive literature in media studies assessing how topics are portrayed in the media. Often this falls under frame analysis (Entman 1993). Frame analysis is a comprehensive, multifaceted qualitative analysis, which can be, but is not always, supported by computational linguistics tools (Schäfer and O'Neill 2017). One aspect of a frame based analysis can be a focusing specifically on word choices used to discuss, or to frame, a topic. Computational linguistics, which combines computational tools with linguistic structure expertise, often focus on the use of word choice specifically, and the relations between words used. This can include focusing on word co-locations (words that appear side by side) (Mertens, De Coninck, and d'Haenens 2022; Alcántara-Plá and Ruiz-Sánchez 2017), and counting of words that are associated with different framings (Mertens, David De Coninck, and Leen d'Haenens).

Computational methods for sentiment classification of words related to the topic of interest in news text has also been explored. However, limitations have been identified, namely, that the framing language used is often subtle due to the expected tone in major media outlets and there is a lack of relevant, news-focused, datasets Hamborg (2020). Additionally, sentiment classification methods focus heavily on positive / negative polarities, which fails to capture the complexity and context-dependency of labeling content in news media as positive or negative (Hamborg 2020).

Methods for sentiment classification are often lexicon based, meaning that a (large) predetermined list of words falling under the sentiments of interest are utilized to 'score' phrases based on whether they contain those words. Limitations of lexicon based methods in general include limited coverage and the need to keep them updated to represent changing language - this is particularly true when working with topics related to rapidly evolving social phenomena, such as gender stereotypes (Cryan et al. 2020). Here we propose an approach that does not require an extensive lexicon, but rather utilizes the feature of contextualized word embeddings (discussed further in section Contextualized Word Embeddings) that the distance between embeddings of different words is related to their level of similarity (Nair, Srinivasan, and Meylan 2020; Wiedemann et al. 2019).

Background

Linguistic Features and Dependency Parsing Dependency parsing is a process used in natural language processing (NLP) to identify the relationships between words in natural language text, and formalize them in a computationally readable way. The structure of a sentence is described in terms of a tree structure, where each word, except the root, has a head, where a head word is the central organizing word of its dependents. A head and its children make up one arc in a given sentence. In this paper dependency parsing is done using spaCy [1], an open-source Python library for natural language processing (NLP). An example of the information learned from dependency parsing a sentence with spaCy is shown in figure 1.

Token	POS	Head	Children
Winner	PROPN	wrote	[]
wrote	VERB	wrote	[Winner, had]
that	SCONJ	had	[]
she	PRON	had	[]
had	VERB	wrote	[that, she, meeting]
a	DET	meeting	[]
30	NUM	minute	[]
-	PUNCT	minute	[]
minute	NOUN	meeting	[30, -]
private	ADJ	meeting	[]
meeting	NOUN	had	[a, minute, private]

Figure 1: An example of dependency parsing information for the phrase *"Winner wrote that she had a 30-minute private meeting."*

Contextualized Word Embeddings A word embedding is a representation of a word as a dense numerical vector where embeddings of similar words are closer together in multidimensional space than dis-similar words. These embeddings are created through training on large amounts of text. Original word embedding models such as gLove (Pennington, Socher, and Manning 2014) and Word2Vec (Mikolov et al. 2013) provide a single embedding per word, whereas large language models (LLMs) (e.g. BERT, ELMo) are capable of generating contextualized word embeddings. These are embeddings where the vector representation of a word depends on the words around it, or the sentence - so there is not just single representation for a given word. Contextualized word embeddings can be generated for full sentences, again with the feature that sentences of similar meaning are closer in distance. In this paper, we generate sentence embeddings using Sentence Transformers (Reimers and Gurevych 2019), a Python framework which uses a pretrained BERT model to create embeddings. The similarity

[1] https://spacy.io/

between word embeddings is calculated by the cosine similarity, or the angle between two vectors, calculated as follows:

$$\cos(\mathbf{a}, \mathbf{b}) = \frac{\sum_{i=1}^{n} \mathbf{a}_i \mathbf{b}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{a}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{b}_i)^2}} \quad (1)$$

Method

We start with explaining the proposed method in general, and then provide the specific details of an experiment to test the method (section Experiment: newsMTSC Dataset) and for a case study utilizing the method on real-world data (section Experiment: BBC News Corpora).

Concept representation. The purpose of our proposed method is to identify the framing of a specific concept within a given corpora of text. What kind of concept can be selected is very flexible, it can, for example, be a single individual, a place, or a specific topic. A textual representation of that target concept needs to be provided; this can be a single word or name, a phrase, or a series of words or phrases, that are used to refer to the concept in the text. From this starting representation of the concept, we are able to identify further mentions of it in the text that do not match the representation. For example, if the target is a person, identified by name, there may be references to the person in the text that do not use their name, such as pronouns. In NLP, the task for identifying theses mentions is coreference resolution. To conduct coreference resolution, we utilize an existing Python module, FastCoref (Otmazgin, Cattan, and Goldberg 2022). Coreference resolution identifies clusters of references, with each cluster referring to a different subject of the text. We then identify the clusters that refer to the target concept by selecting the clusters that contain the word, or phrase, we have selected to represent the concept.

Relevant words. Once we have identified all mentions of the target concept in the text, we are able to identify the words that are used in relation to the target concept. We call these the *relevant words*. To do this we utilize dependency parsing (described in section Linguistic Features and Dependency Parsing) to identify the relationship between the words in the text. We utilize the parsed dependency tree to identify these relevant words. Our definition of relevant words can be summarized as all adjectives, nouns and verbs that are either the head of the target word, as well as the heads of any coreference of the target, along with all other children of those heads. In this way we identify the arcs within the text which directly relate to mentions of the target. In the example shown in figure 1, if the subject is 'Winner' then the coreferences are 'Winner' and 'she' and the words contained in the relevant arcs are *wrote, had, that, meeting*. In order to focus on the sentiment conveying words, only the adjectives, verbs and nouns are selected as the final relevant words; in the case of our example, the relevant words are *wrote, that, meeting*. These relevant words are then represented as contextualized word embeddings, as described in the section Contextualized Word Embeddings.

The measurement framework. Our method utilizes measures of similarities between contextual word embeddings to assess the framing of the target concept. To this end, a framework by which to measure it must be selected, and the measurement framework must be represented as text, and thus as contextual word embeddings. In this work we propose three different measurement frameworks. The first, which is used in the experiments in both section Experiment: newsMTSC Dataset and Experiment: BBC News Corpora, is a measure of positive and negative sentiment. The positive sentiment is represented by the contextualized word embedding of the following:

> *admire, amazing, assure, celebration, charm, eager, enthusiastic, excellent, fancy, fantastic, frolic, graceful, happy, joy, luck, majesty, mercy, nice, patience, perfect, proud, rejoice, relief, respect, satisfactorily, sensational, super, terrific, thank, vivid, wise, wonderful, zest.*

The negative sentiment is represented by a contextualized embedding of the following:

> *abominable, anger, anxious, bad, catastrophe, cheap, complaint, condescending, deceit, defective, disappointment, embarrass, fake, fear, filthy, fool, guilt, hate, idiot, inflict, lazy, miserable, mourn, nervous, objection, pest, plot, reject, scream, silly, terrible, unfriendly, vile, wicked.*

This measurement framework is based on words which are common among three commonly used sentiment lexicons in Jurafsky and Martin (2023). We use these short lists of words, rather than words from an entire lexicon so that the framework is readable, understandable and maintainable. Embeddings of a small amount of words for representing a specific framework have been utilized by (Caliskan, Bryson, and Narayanan 2017), who found that a small number of select words reduced the amount of noise created by the multiple meanings of many words.

We utilize two frameworks from (Caliskan, Bryson, and Narayanan 2017) as two additional frameworks in the experiment in section Experiment: BBC News Corpora. In that work, both of these frameworks were used to represent existing gender-based stereotypes that have been previously identified in social science literature. Namely, the association of females with family, more than career (compared to males) and with arts, more than sciences (compared to males). The words used by Caliskan, Bryson, and Narayanan (2017), as well as our approach, to represent the four poles of family, career, arts and science are as follows:

> *Family*: home, parents, children, family, cousins, marriage, wedding, relatives
> *Career*: executive, management, professional, corporation, salary, office, business, career
> *Arts*: poetry, art, Shakespeare, dance, literature, novel, symphony, drama
> *Science*: science, technology, physics, chemistry, Einstein, NASA, experiment, astronomy

The framework nodes are represented as contextualized word embeddings, as described in described in section Contextualized Word Embeddings, and the cosine similarity between these and the embeddings of the relevant words are then calculated. We are then able to compare distances (or similarity) of the framing of various concepts to the framework nodes, thus providing a comparative measure of the framing of the concepts.

Experiment: newsMTSC Dataset

Method: newsMTSC Dataset

In order to validate the effectiveness of our method of comparing word embeddings of relevant text to embeddings representing a framework, we tested it on a dataset of labeled text snippets. The NewsMTSC dataset (Hamborg and Donnay 2021) consists of human-labeled sentences, which were sampled from news articles from online US news outlets; it was developed for the purposes of target-dependent sentiment classification (TSC) in news articles. Each sentence is labeled for whether it conveys positive, negative or neutral sentiment for a specific target, namely a person mentioned in the sentence. Labels are derived from an aggregation of the labels from multiple human annotators.

We use 4163 labeled sentences containing a single target as the text corpora, the target portrayal is labeled as positive in 1815 sentences and as negative in 2348 sentences. The dataset includes the name of the target for each sentence as well as character indexes for that name in the sentence; this is our textual representation of the target concept. We then complete the coreference resolution and the selection of relevant words using the method described above. This leaves us with a list of relevant words per labeled sentence. We then create two aggregated lists, one of all relevant words used in relation to targets who were portrayed positively and another for targets who were portrayed negatively. These lists are then used to create a positive and a negative sentence embedding, which are created from one text string containing all of the positive or negative relevant words, separated by commas.

As a baseline for comparison, we create sentences intended to represent a text containing a mix of positive and negative sentiment towards multiple targets. We do this by concantenating two sentences from the newsMTSC dataset, one with a positive label, and one with a negative label. We created 120 of these sentences and refer to them as neutral sentences. The same process was conducted with them as for the positive and negative sentences.

The framework utilized is the positive / negative sentiment framework described in Method section. The input strings used to create the embeddings is again a list of each word in the framework node.

To calculate a distribution of the similarity between the two concept word embeddings and the two framework nodes, we conduct bootstrapping of the cosine similarity at 1000 iterations. This is done by sampling, with replacement, from the list of words representing the two concepts. We are then able to compare the mean similarity of the positive and negative concept word embeddings with the positive and negative framework nodes. We expect that the relevant words from the positive concept word embeddings will be closer to the positive framework nodes, and vice-versa. If this is the case, then our method is consistent with the results of human labeling and we consider it as validation that our method can effectively identify whether a target in a text corpus is being framed in a specific manner. Results from this experiment are reported in the following subsection.

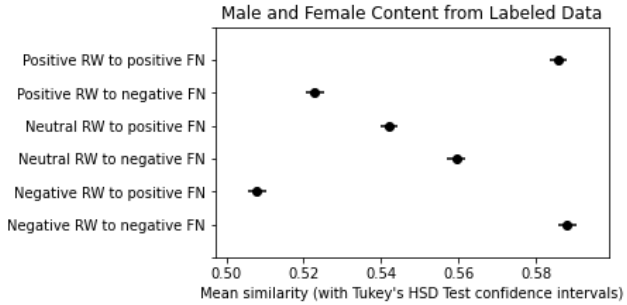

Figure 2: Results of the newMTSC experiment. Presented here are the similarities between embeddings of Relevant Words (RW) identified in the positively and negatively labeled sentences and the Framework Nodes (FN) of the positive / negative framework.

Results: newsMTSC Dataset

The similarities of each of the three concept groups (positive, negative and neutral) to the two framework nodes (positive and negative) is measured, resulting in six similarities to be measured. A one-way ANOVA test was performed to compare the resultant six mean similarities. The one-way ANOVA showed that there is a statistically significant difference between the groups (F=834.917, p=0.000). In order to see which of the groups were significantly different from each other we perform post-hoc testing in the form of the Tukey Honest Statistical Difference (Tukey HSD) test. The results of the test are presented in table 1. We find that all means similarities are significantly different from all others, except the difference between the the negative relevant words to the negative framework node and the positive relevant words to positive framework node.

Specifically, we find that the positive relevant words have a higher cosine similarity to the positive framework node than to the negative node embedding, while negative relevant words have a higher cosine similarity to the negative framework node than to the positive framework node, which can be seen visually in figure 2. This result is in line with what is expected if our proposed method is successfully identifying positive or negative sentiment towards a target in text. Additionally, we see that the neutral relevant words are closer to the negative node than to the positive node, however, the difference in similarity, to both nodes, is much smaller than in the case of the positive and negative relevant words.

Experiment: BBC News Corpora

Using the approach on real world data.

Table 1: Results of the Tukey HSD test here of the similarities between embeddings of Relevant Words (RW) identified in the positively and negatively labeled sentences and the Framework Nodes (FN) of the positive / negative framework in the newsMTSC dataset.

Group 1	Group 2	Mean Diff.	p-value	95% CI
Negative RW to negative FN	Negative RW to positive FN	-0.0801	-0.0**	(-0.0847) - (-0.0755)
Negative RW to negative FN	Neutral RW to negative FN	-0.0286	-0.0**	(-0.0332) - (-0.024)
Negative RW to negative FN	Neutral RW to positive FN	-0.046	-0.0**	(-0.0505) - (-0.0414)
Negative RW to negative FN	Positive RW to negative FN	-0.0652	-0.0**	(-0.0698) - (-0.0607)
Negative RW to negative FN	Positive RW to positive FN	-0.0023	0.7093	(-0.0069) - 0.0023
Negative RW to positive FN	Neutral RW to negative FN	0.0515	-0.0**	0.0469 - 0.0561
Negative RW to positive FN	Neutral RW to positive FN	0.0342	-0.0**	0.0296 - 0.0387
Negative RW to positive FN	Positive RW to negative FN	0.0149	-0.0**	0.0103 - 0.0195
Negative RW to positive FN	Positive RW to positive FN	0.0778	-0.0**	0.0733 - 0.0824
Neutral RW to negative FN	Neutral RW to positive FN	-0.0174	-0.0**	(-0.0219) - 0.0128
Neutral RW to negative FN	Positive RW to negative FN	-0.0366	-0.0**	(-0.0412) - (-0.032)
Neutral RW to negative FN	Positive RW to positive FN	0.0263	-0.0**	0.0217 - 0.0309
Neutral RW to positive FN	Positive RW to negative FN	-0.0193	-0.0**	(-0.0238) - (-0.0147)
Neutral RW to positive FN	Positive RW to positive FN	0.0437	-0.0**	0.0391 - 0.0483
Positive RW to negative FN	Positive RW to positive FN	0.0629	-0.0**	0.0584 - 0.0675

Notes:
** Significant at $p < 0.01$
* Significant at $p < 0.05$

Method: BBC News Corpora

We applied our approach to real world data, namely, a corpus of english language BBC News stories published between January 1, 2023 and August 1, 2023. All analysis was done on the full text of the bodies of the news stories. We apply our analysis to the question of the framing of males and females in the BBC News stories written on the topic of UK politics. We identified 1036 news stories on UK politics through the use of content tags.

In order to identify mentions of males and females in the text we conduct coreference resolution, as described in section Method. We then identify all coreference clusters that contain male or female pronouns and separate those clusters into two groups, male and female. We then select all relevant words related to those concepts, as described above. This leaves us with a list of relevant words per gender per news story. For each news story, we are then able to create a male relevant words embedding and a female relevant words embedding from a text string consisting of the corresponding list of words. We compare these embeddings to each of the three frameworks described above: positive / negative, science / arts and career / family.

We identify, per news story, an average of 4.0 mentions of females (with 11.6 relevant words) and 15.93 mentions of males (with 27.3 relevant words). For this reason, we do not need to use bootstrapping here to calculate the distributions of the similarities between the relevant words and the framework nodes as we are able to calculate the cosine similarities per news story. Results from this experiment are reported in the following subsection.

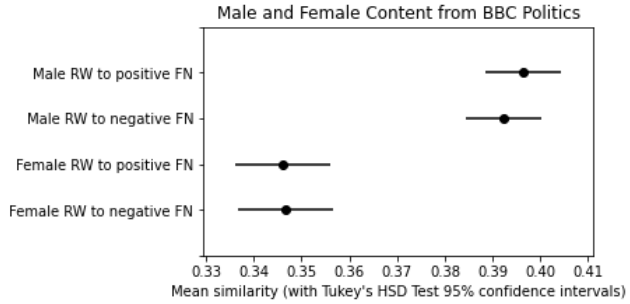

Figure 3: Results based on the BBC News Corpora of UK politics tagged news stories. Presented here are the similarities between embeddings of Relevant Words (RW) used in relation to male and female subjects and the Framework Nodes (FN) of the positive / negative framework.

Results: BBC News Corpora

Positive / negative framework The similarities of each of the two relevant word groups (male and female) to the two framework nodes (positive and negative) is measured, resulting in four similarity means. A one-way ANOVA test comparing the four measures showed that there is a statistically significant difference between the groups (F=32.018, p=0.000). According to the post-hoc Tukey HSD the mean similarities of the female relevant words to both positive and negative nodes are significantly lower than the male relevant words are to both nodes (full results in table (2). However, there is no significant difference between similarities to the positive and negative nodes within the genders. This suggests that the male relevant words in the BBC politics stories

Table 2: Results of the Tukey HSD test of the similarities between embeddings of Relevant Words (RW) used in relation to male and female subjects in BBC UK politics news stories to the Framework Nodes (FN) of the positive / negative framework.

Group 1	Group 2	Mean Diff.	p-value	95% CI
Female RW to negative FN	Female RW to positive FN	-0.0005	0.9999	(-0.0202) - 0.0193
Female RW to negative FN	Male RW to negative FN	0.0458	0.0 **	0.028 - 0.0637
Female RW to negative FN	Male RW to positive FN	0.0498	0.0 **	0.0319 - 0.0677
Female RW to positive FN	Male RW to negative FN	0.0463	0.0**	0.0284 - 0.0641
Female RW to positive FN	Male RW to positive FN	0.0503	0.0**	0.0324 - 0.0681
Male RW to negative FN	Male RW to positive FN	0.004	0.9154	(-0.0118) - 0.0198

Notes:
** Significant at the $p < 0.01$
* Significant at the $p < 0.05$

are both more positive and negative than the female relevant words, which could mean that less sentiment filled language is used when discussing female subjects than males.

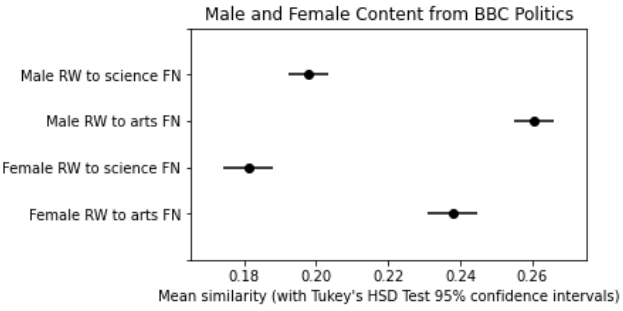

Figure 4: Results based on the BBC News Corpora of UK politics tagged news stories. Presented here are the similarities between embeddings of Relevant Words (RW) used in relation to male and female subjects and the Framework Nodes (FN) of the arts / science framework.

Art / science framework The similarities of each of the two relevant word groups (male and female) to the two framework nodes (art and science) is measured, resulting in four similarity means. A one-way ANOVA test comparing the four means showed that there is a statistically significant difference between the groups (F=118.605, p=0.000). According to the post-hoc Tukey HSD each of the mean similarities are significantly different from each of the others. The results can be seen in table 3. As can be seen in figure 4, male relevant words are closer to both the arts and science nodes.

Career / family framework The comparison of the similarities of each of the two relevant word groups (male and female) to the two framework nodes (career and family) results in four similarity means. A one-way ANOVA test comparing the four means showed that there is a statistically significant difference between the groups (F=181.674, p=0.000). According to the post-hoc Tukey HSD each of the mean similarities are significantly different from each of the others, except for the mean similarities between male and female relevant words to the family framework node, as can be seen in table 3. In figure 5 we see that male relevant words are closer to the career node than female relevant words, however, both male and female relevant words are similar distances (not significantly different) from the family node.

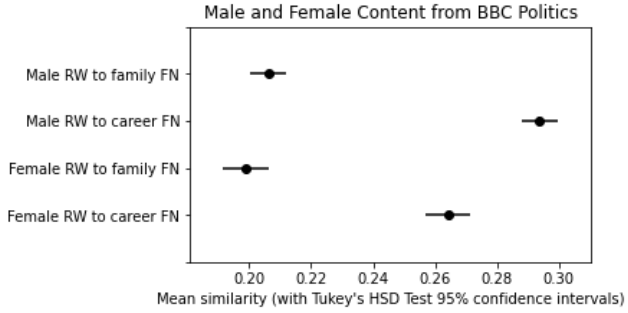

Figure 5: Results based on the BBC News Corpora of UK politics tagged news stories. Presented here are the similarities between embeddings of Relevant Words (RW) used in relation to male and female subjects and the Framework Nodes (FN) of the family / career framework.

Discussion We see a near-consistent effect whereby semantic similarities between topic-referent words and male-referent words and are greater than those to female-referent words—this holds true for all but the family topic. This may be an indication that more sentiment-filled and topic specific language is used around male-referents compared to female-referents but it does not support the presence of any of the tested gender-based stereotypes in the BBC politics dataset. In the following section, we discuss the further work needed to allow additional interpretations of these findings and to parse the various factors influencing embedding differences

Limitations, Future Work and Conclusion

Limitations of the work include that it relies entirely on words in text; clearly additional factors like news story

Table 3: Results of the Tukey HSD test of the similarities between embeddings of Relevant Words (RW) used in relation to male and female subjects in BBC UK politics news stories to the Framework Nodes (FN) of the arts / sciences framework.

Group 1	Group 2	Mean Diff.	p-value	95% CI
Female RW to arts FN	Female RW to science FN	-0.0566	0.0**	(-0.0704) - (-0.0428)
Female RW to arts FN	Male RW to arts FN	0.0225	0.0**	0.01 - 0.0349
Female RW to arts FN	Male RW to science FN	-0.0399	0.0**	(-0.0524) - (-0.0274)
Female RW to science FN	Male RW to arts FN	0.0791	0.0**	0.0666 - 0.0915
Female RW to science FN	Male RW to science FN	0.0167	0.0033**	0.0042 - 0.0292
Male RW to arts FN	Male RW to science FN	-0.0624	0.0**	(-0.0734) - (-0.0514)

Notes:
** Significant at the $p < 0.01$
* Significant at the $p < 0.05$

Table 4: Results of the Tukey HSD test of the similarities between embeddings of Relevant Words (RW) used in relation to male and female subjects in BBC UK politics news stories to the Framework Nodes (FN) of the career / family framework.

Group 1	Group 2	Mean Diff.	p-value	95% CI
Female RW to career FN	Female RW to family FN	-0.0654	0.0**	-0.0798 -0.051
Female RW to career FN	Male RW to career FN	0.0296	0.0**	0.0165 0.0426
Female RW to career FN	Male RW to family FN	-0.058	0.0 **	-0.0711 -0.045
Female RW to family FN	Male RW to career FN	0.095	0.0**	0.0819 0.108
Female RW to family FN	Male RW to family FN	0.0073	0.469	-0.0057 0.0204
Male RW to career FN	Male RW to family FN	-0.0876	0.0**	-0.0991 -0.0761

Notes:
** Significant at the $p < 0.01$
* Significant at the $p < 0.05$

placement or images are not included in the analysis. Further, even in the text, some factors may not be identified. For example, we see in the MTSC news dataset the following sentence which uses punctuation to convey sarcasm: *Hillary Clinton takes "responsibility" for loss, but says others contributed*. Additionally, while we test the positive / negative framework against labeled text data, we do not have an analogous labeled dataset to test the family / career or art / science frameworks. Creation of labeled datasets for frameworks of interest needs to be part of future work.

It is further important to conduct further work to understand how we can interpret the similarities we are examining in this method. Firstly, it is important to better understand the impact of potential confounding factors, such as the differences in the amount of mentions of different topics and of the amount of relevant words found. It is also important to parse out the potential impact of word embedding bias (Caliskan, Bryson, and Narayanan 2017) - this could include words that have a difference in embedding similarity in relation to the framework (such as more negative or more positive), but are not interpreted as such by human readers. In the case that such words are used more in conjunction with one topic than another, this would bias the results. Secondly it is important to identify the best format for the textual representation of the relevant words and the framework nodes. Specifically, we currently create a list of words to be transformed into sentence embeddings, however as BERT is trained on natural language sentences, the list of words may not be the most suitable representation. Finally, the method for identifying relevant words can be further updated with linguistic expertise; the current approach, where word dependencies of all input sentences are identified, allows for transparent explanation of how the words are chosen as well as flexibility to improve how that is done.

In terms of use for monitoring news content, the baselines for similarity comparisons, as well as what changes in similarities trigger further investigation, needs to be defined in conjunction with editorial experts. In reference to the current results, it remains to be seen what the embedding distances we found mean in practice, and thus how they should be interpreted. Conducting similar analysis with other text corpora, including from other news providers, and with human-labeled data is required for further understanding. All further work should continue to consider the existing work in media studies and framing analysis, as well incorporate feedback from news editors, to determine best design decisions and explanations of output to support the work of experts in those areas.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation program under Marie Sklodowska-Curie Actions (grant agreement number 860630) for the project "NoBIAS - Artificial Intelligence

without Bias". This work reflects only the authors' views and the European Research Executive Agency (REA) is not responsible for any use that may be made of the information it contains.

References

2016. An Agreement Between Her Majesty's Secretary of State for Culture, Media and Sport and the British Broadcasting Corporation.

Ada Lovelace Institute. 2022. Inform, educate, entertain... and recommend? Exploring the use and ethics of recommendation systems in public service media. Technical report.

Alcántara-Plá, M., and Ruiz-Sánchez, A. 2017. The framing of Muslims on the Spanish Internet. *Lodz Papers in Pragmatics* 13(2):261–283. Num Pages: 261-283 Place: Lodz, Germany Publisher: Walter de Gruyter GmbH.

Boididou, C.; Sheng, D.; Mercer Moss, F. J.; and Piscopo, A. 2021. Building Public Service Recommenders: Logbook of a Journey. In *Fifteenth ACM Conference on Recommender Systems*, 538–540. Amsterdam Netherlands: ACM.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334):183–186. Publisher: American Association for the Advancement of Science Section: Reports.

Cryan, J.; Tang, S.; Zhang, X.; Metzger, M.; Zheng, H.; and Zhao, B. Y. 2020. Detecting Gender Stereotypes: Lexicon vs. Supervised Learning Methods. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–11. Honolulu HI USA: ACM.

Entman, R. M. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of communication* 43(4):51–58.

Hamborg, F., and Donnay, K. 2021. NewsMTSC: A Dataset for (Multi-)Target-dependent Sentiment Classification in Political News Articles. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1663–1675. Online: Association for Computational Linguistics.

Hamborg, F. 2020. Media Bias, the Social Sciences, and NLP: Automating Frame Analyses to Identify Bias by Word Choice and Labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 79–87. Online: Association for Computational Linguistics.

Jurafsky, D., and Martin, J. H. 2023. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall, 3rd edition.

Mertens, S.; David De Coninck; and Leen d'Haenens. A report on legacy media coverage of migrants. Technical report.

Mertens, S.; De Coninck, D.; and d'Haenens, L. 2022. The Twitter debate on immigration in Austria, Germany, Hungary, and Italy: Politicians' articulations of the discourses of openness and closure.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nair, S.; Srinivasan, M.; and Meylan, S. 2020. Contextualized word embeddings encode aspects of human-like word sense knowledge. In Zock, M.; Chersoni, E.; Lenci, A.; and Santus, E., eds., *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, 129–141. Online: Association for Computational Linguistics.

Otmazgin, S.; Cattan, A.; and Goldberg, Y. 2022. F-coref: Fast, Accurate and Easy to Use Coreference Resolution. In Buntine, W., and Liakata, M., eds., *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: System Demonstrations*, 48–56. Taipei, Taiwan: Association for Computational Linguistics.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Piscopo, A.; Panteli, M.; and Penna, D. 2019. Data-Driven Recommendations in a Public Service Organisation. In *Proceedings of the 23rd International Workshop on Personalization and Recommendation on the Web and Beyond*, ABIS '19, 23–24. New York, NY, USA: Association for Computing Machinery.

Reimers, N., and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *EMNLP/IJCNLP (1)*, 3980–3990. Association for Computational Linguistics.

Schäfer, M., and O'Neill, S. 2017. Frame Analysis in Climate Change Communication.

Wiedemann, G.; Remus, S.; Chawla, A.; and Biemann, C. 2019. Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. In *Proceedings of the 15th conference on natural language processing, KONVENS 2019, erlangen, germany, october 9-11, 2019*. tex.citedby: 0 tex.cites: 0.