SCHOOL OF ECONOMICS AND MANAGEMENT

# Do Words Matter?

## Gendered Language and Sentiment Polarity in U.S. News Headlines Across #MeToo and COVID-19


Sophie Ordish

Department of Economics

Lund University School of Economics and Management



**Supervisors: Eva Ranehill & Linda Erwe**

Data Analytics and Business Economics: Master Essay 15 ECTS

Seminar date: June, 2025

# CONTENTS

# Do Words Matter?*
# Gendered Language and Sentiment Polarity in U.S. News Headlines Across #MeToo and COVID-19

Sophie Ordish†

June 2025

*Acknowledgements: First and foremost, I would like to express my deepest gratitude to my supervisors, Eva Ranehill and Linda Erwe, for their invaluable guidance, encouragement, and expertise throughout this thesis process. Their thoughtful feedback and unwavering support helped shape this project and pushed me to think critically and rigorously at every stage.

I am also incredibly grateful have studied with DABEst cohort. This unforgettable year has been shaped by your friendship, and your good humour has proven to be a constant source of entertainment. Without your moral support and much-needed motivation I would not be where I am today. In no particular order, thank you: Alina, Analiz, Dilay, Denisse (for fuelling us with home-baked goods), Elide (my unofficial thesis partner), Helga, Patricie, and Viktorija. Thank you for the countless hours of discussion and collaboration. Special thanks go to those who reviewed drafts, shared feedback, or simply encouraged me to take a break when it was needed most.

Thank you Beddingfield (2004) for reminding me to never fear the blank page before me.

Finally, I am eternally grateful to my family for their constant support, love, and belief in me. Their encouragement has been a steady foundation throughout my academic journey. And to Elvire, for her kindness and support during the final stretch of this thesis. None of this would have been possible without all of you.

†Lund University School of Economics and Management

**Abstract**

This thesis investigates how women are portrayed in U.S. news headlines through the lens of gender bias and sentiment polarity. Using a dataset of over 130,000 headlines from major American news outlets between 2005 and 2021, the study combines a curated dictionary-based approach to detect stereotypical language with transformer-based sentiment analysis to measure tone. Two major societal disruptions, the #MeToo movement and the COVID-19 pandemic, are examined as focal events for structural change in media language. Time series models, including changepoint detection and ARMA forecasting, are employed to test for statistically significant shifts in bias and polarity. The results show that while #MeToo increased the visibility of women's stories, it had limited impact on the underlying tone or bias of media coverage. In contrast, the COVID-19 pandemic introduced greater volatility and emotional intensity in headlines, reflecting more pronounced editorial shifts. These findings suggest that despite increased representation, the linguistic framing of women remains persistently stereotyped and emotionally narrow, limited to conventional portrayals such as caregiving, vulnerability, or victimhood, with little room for agency, anger, or authority. The thesis contributes to literature on feminist media studies, computational linguistics, and event inference by offering a replicable framework for measuring media bias over time. Future research could examine how women in leadership or professional roles are portrayed in business and economic news, exploring shifts in bias around female CEOs, politicians, or entrepreneurs. This could reveal how media framing influences perceptions of competence and authority, and whether it correlates with real-world outcomes like hiring trends or board representation. Further work could also extend the analysis to social media platforms to build a more comprehensive view of gender representation in digital media.
**Keywords:** Gender bias, news media, sentiment analysis, time series, computational linguistics

# 1 INTRODUCTION

Women remain considerably underrepresented in mainstream news media. When they do appear, their portrayal is often shaped by gendered or stereotypical language that reinforces traditional roles and entrenched biases (Macharia, 2021; Santoniccolo et al., 2023). These linguistic patterns play a powerful role in shaping public perceptions of competence, value, and agency. Headlines, in particular, are a highly influential form of journalistic communication. Designed to capture attention and prioritise relevance, they not only summarise news stories but also frame their interpretation (Dor, 2003). In today's digital media landscape, many readers encounter only the headline of a story, often via social media or notifications (Boczkowski and Mitchelstein, 2013; Reuters Institute for the Study of Journalism, 2023), and as such, headlines carry disproportionate weight in shaping public opinion and framing social discourse.

This thesis investigates how gendered and stereotypical language in headlines about women has evolved over time. It focuses on the United States, home to some of the most globally influential media organisations, where news outlets exert significant power over both the prominence of issues and how they are interpreted (McCombs and Shaw, 1972; Entman, 1993). The aim is to identify whether major social disruptions have coincided with shifts in the linguistic portrayal of women. Two events are of particular interest: the #MeToo movement and the COVID-19 pandemic. These moments differ in cause and character, yet both restructured public discourse, potentially altering how women were framed in the media.

To address this, the study poses the following research question:

*By analysing the presence of gendered and/or stereotypical language in women's news headlines, are there identifiable shifts in bias or sentiment polarity surrounding major social events such as #MeToo or the COVID-19 pandemic?*

The null hypothesis assumes linguistic framing remained stable; the alternative posits significant structural change.

The methodological approach constructs two daily time series from a corpus of U.S. headlines between 2005 and 2021. The first captures gender bias using a curated dictionary adapted from Xu et al. (2019). The second quantifies sentiment polarity, whether the framing is positive, negative, or neutral, using a transformer-based classifier (Liu et al., 2019). Both series are analysed through a hybrid strategy combining unsupervised

changepoint detection and event-based ARMA forecasting.

The #MeToo movement, which gained global momentum in late 2017 following a viral tweet by Alyssa Milano[1] brought issues of sexual harassment, gender-based violence, and structural inequality to the centre of public and media attention. It challenged dominant narratives surrounding gender, credibility, and institutional power, especially in the workplace (Mendes et al., 2018; Fileborn and Loney-Howes, 2019). As such, #MeToo offers a compelling case for examining whether increased visibility for women also translated into shifts in the emotional tone and linguistic bias of media representations. Adding to the findings of Starkey et al. (2019), who showed that early media coverage of the movement often focused on individual narratives and celebrity cases rather than systemic critique, this study tests whether such framing patterns were accompanied by measurable shifts in sentiment polarity and gender bias at the headline level.

By contrast, the COVID-19 pandemic, declared by the World Health Organization (2020) on March 11, 2020, reshaped nearly every facet of public and private life. Research shows that women bore the brunt of the pandemic's social and economic fallout, taking on more unpaid care work, facing job losses in feminised sectors, and working disproportionately in frontline roles (Wenham et al., 2020; Cowie et al., 2021). Despite these unequal burdens, media narratives during this period often reinforced traditional gender roles, portraying women primarily as caregivers or victims. As Gill and Orgad (2022) argue, seemingly benign media framings that emphasised calm, self-care, and positivity during the crisis tended to mute anger and obscure structural critiques, ultimately reinforcing neoliberal assumptions and gendered inequalities (Seluman et al., 2024). COVID-19 thus serves as a contrasting disruption to #MeToo, one not driven by gender politics but nonetheless marked by deeply gendered consequences. Both events offer distinct yet complementary lenses through which to assess whether and how the emotional and ideological framing of women in news headlines shifted during periods of social upheaval.

This study finds that the #MeToo movement led to only marginal shifts in the linguistic patterns of bias and sentiment polarity in headlines about women. In contrast, the COVID-19 pandemic produced sharper and more persistent fluctuations in both sentiment polarity and its volatility. These fluctuations appear less tied to gender politics

[1] *Tweet*: "If you've been sexually harassed or assaulted write 'me too' as a response to this tweet." *Retweet*: "Me too. Suggested by a friend: 'If all the women who have been sexually harassed or assaulted wrote 'Me too.' as a status, we might give people a sense of the magnitude of the problem.'"

than to broader conventions of crisis reporting, which tend to favour emotionally charged language. The contrast between these two events suggests that while visibility for women may increase during moments of public reckoning, the underlying editorial norms and cultural scripts that shape their portrayal remain strikingly resilient. These findings have important implications for understanding the limits of media reform. Whilst significant social events may temporarily amplify women's presence in the news, they do not necessarily transform the emotional or ideological framing through which they are represented.

This research contributes to four intersecting literatures: media framing, feminist linguistics, computational text analysis, and event inference in time series. By integrating machine learning with social theory, the study reveals how empirical methods can uncover subtle but consequential patterns in public discourse. It highlights the persistence of gendered framings in headlines and suggests that meaningful change may require not only increased representation, but also a rethinking of the linguistic and emotional framing of women in the news.

The remainder of this thesis is structured as follows: Section 2 reviews the relevant literature on gender bias in media, sentiment analysis, and event-based causal inference. Section 3 describes the data sources and pre-processing procedures. Section 4 outlines the empirical strategy, including the construction of bias and sentiment measures and the application of time series models. Section 5 presents the main findings. Section 6 discusses the study's limitations and proposes directions for future research. Finally, Section 7 concludes with a summary of the contributions and implications of the study.

# 2 LITERATURE REVIEW

This literature review draws on four strands of research relevant to the study of gendered language in U.S. media headlines. First, it examines how women are historically represented in the news, with particular attention to stereotypical portrayals and underrepresentation. Second, it explores the broader societal and technological consequences of media bias, especially in the context of machine learning and artificial intelligence. Third, it reviews methodological approaches for detecting bias and sentiment in text corpora using natural language processing (NLP). Finally, it considers statistical techniques for detecting structural change over time, with a focus on event-based and time series models.

Together, these perspectives inform the analytical framework of this thesis, which investigates how gendered and stereotypical language in headlines has evolved in response to key social disruptions, namely the emergence of the #MeToo movement and the onset of the COVID-19 pandemic.

## *2.1 Gender Representation in Headlines*

Decades of research demonstrate that women remain underrepresented in news media and, when present, are frequently framed in stereotypical terms. The Global Media Monitoring Project (GMMP), a large-scale longitudinal analysis, found that women comprised only 25% of the individuals featured in global news stories in 2020, up only modestly from 17% in 1995 (Macharia, 2021).

Computational studies echo these findings. Santoniccolo et al. (2023), for example, analysed headlines from several Western countries and found that women were not only less frequently mentioned than men, but were more likely to be described using stereotypical and relational language (e.g. appearance, family roles, emotionality) rather than associated with power or expertise. Similarly, Dragaš (2012) used critical discourse analysis to show how gender norms in Serbian newspapers reinforce unequal portrayals of women, even in otherwise progressive contexts.

Beyond news media, Xu et al. (2019) identify narrative gender biases in broader cultural media, such as books and films, using word embeddings to trace emotional arcs. They show that female characters' happiness often depends on male counterparts, rein-

forcing traditional gender scripts. Their framework, especially the use of custom dictionaries and embedding-based measures of emotion, has informed the methodological design of this thesis.

Cultural beliefs also persist despite structural change. Haines et al. (2016) find that traditional gender stereotypes remain widely endorsed in the United States, even as women's workforce participation increases. This disconnect between social progress and cultural representation helps explain the durability of biased language in news discourse.

### *2.2 Societal and Technological Consequences of Media Bias*

Bias in media texts has ramifications beyond journalism. As digital systems increasingly rely on large text corpora, including news headlines, for training, media bias becomes embedded in machine learning models. Hitti et al. (2019) show that natural language inference models trained on biased data reproduce gender stereotypes in textual reasoning tasks, whilst Lu et al. (2020) find that large language models (LLMs), such as GPT-3, systematically associate male pronouns with high-status occupations and female pronouns with caregiving roles, even under neutral prompts.

These patterns are not limited to lexical choice. Fang et al. (2023) demonstrate that gender and racial biases also emerge at syntactic and semantic levels, affecting narrative structure and credibility. Such findings raise concerns about the feedback loop between biased human-generated content and the AI systems trained on it.

At the same time, many tools used to measure bias, such as LLMs, are trained on these same corpora, raising challenges in knowledge acquisition. If measurement tools reproduce the stereotypes they aim to detect, then bias identification must be complemented by interpretive and context-sensitive methods. This study contributes to this effort by combining lexicon-based and model-based tools within a bounded dataset of mainstream U.S. news headlines, selected for its social relevance to public discourse on gender.

### *2.3 Detecting Gender Bias and Sentiment in Textual Data*

NLP methods have enabled increasingly sophisticated approaches to detecting gender bias and sentiment in large-scale textual corpora. Lexicon-based approaches remain foundational, allowing researchers to track stereotypical language over time with interpretable and transparent metrics. For example, Hu and Kearney (2021) analyse collocations be-

tween gendered nouns and descriptive terms to reveal framing patterns in political discourse. Applied over time-series data, such methods can uncover both continuity and change in cultural representations of gender.

Vector-based methods offer deeper contextual modelling. As discussed above, Xu et al. (2019) use word embeddings and part-of-speech analysis to map emotional dependencies and stereotypical descriptors in fictional narratives. Their approach of linking linguistic structure to narrative framing offers a transferable methodology for studying media texts such as headlines, where characterisation and emotional tone are often compressed into short, high-impact language.

Sentiment analysis complements bias detection by capturing shifts in emotional framing. Traditional lexicon-based models like VADER are popular for their speed and simplicity but are limited in their ability to handle sarcasm, negation, and contextual nuance. Transformer-based models, such as BERT and RoBERTa, significantly improve upon this by leveraging contextual embeddings and bidirectional attention. Purohit et al. (2025) find that BERT outperforms earlier sentiment classifiers and rule-based methods when applied to news articles, producing probability distributions over sentiment classes that allow for continuous polarity scoring. This study adopts a similar transformer-based approach for sentiment analysis.

While recent research increasingly adopts embeddings and deep contextual models to capture nuanced semantic patterns, this thesis employs a lexicon-based approach to measure gender bias. This choice reflects several advantages in the context of time-series media analysis: lexicon methods are transparent, computationally efficient, and less sensitive to model drift or domain adaptation over time. Although dictionary-based methods may not capture all contextual subtleties or emergent forms of language use, they are particularly well-suited for longitudinal studies that prioritise interpretability and theoretical alignment. Dictionaries have long played a foundational role in semantic analysis, from early measures of word similarity (Budanitsky and Hirst, 2006; Jiang and Conrath, 1997) to the post-processing and refinement of word embeddings (Faruqui et al., 2015; Tissier et al., 2017; Bollegala et al., 2016), and more recently, efforts to debias pretrained representations (Glavaš and Vulić, 2018). In media analysis, where transparency and construct validity are critical, lexicon methods offer a robust means of tracing stereotypical gender language over time. This study builds on the approach of Xu et al. (2019), adapting

their curated term sets to the domain of news headlines in order to track gendered and stereotypical language across a 16-year period.

### *2.4 Event Analysis and Causal Inference in Time Series*

While much of the existing research documents gender bias descriptively, few studies evaluate whether major societal events lead to measurable linguistic change. To do so, this thesis adapts methods from event studies and causal inference in time series.

The event study framework, originally developed by MacKinlay (1997) for financial markets, models expected outcomes over a pre-event window and compares them to post-event observations. Deviations are interpreted as the event's effect. Though rooted in economics, this method is increasingly applied to public health and communication studies.

To strengthen causal claims, the analysis also draws on regression discontinuity in time (RDiT) designs. As Lee and Lemieux (2010) explain, RDiT compares outcomes immediately before and after a known threshold under the assumption of local continuity. However, as Hausman and Rapson (2018) caution, standard RDD assumptions often break down with time as the running variable due to autocorrelation and non-stationarity. They recommend integrating ARMA models to account for temporal structure and improve counterfactual estimation, as adopted in this study.

Finally, the thesis takes a theoretically informed view of causality in cultural data. As Gerring (2005) argues, social change is rarely attributable to single events. Instead, it emerges from overlapping conditions, for example, media norms, institutional responses, and audience attitudes, that interact in complex ways. Accordingly, statistical breakpoints are interpreted within broader socio-cultural narratives, rather than treated as mechanically causal.

## 3 DATA

The dataset used in this study consists of news headlines in which women are the primary subjects. The headlines were collected from a range of major English-language news publications and cover the period from 2005 to 2021. The goal of the dataset construction was to assemble a diverse and representative corpus for analysing patterns in language, sentiment polarity, and bias related to the portrayal of women in media over time.

The source of this data is the *Women in Headlines: Bias* dataset from Kaggle (2022), a popular online resource for data, and licensed by MIT. Headlines containing one or more of the following keywords (*women, woman, girl, female, lady, ladies, she, her, herself, aunt, grandmother, mother, sister, daughter, wife, mom, mum, girlfriend, mrs, niece*) were scraped from the top news publications in the United States, United Kingdom, India, and South Africa, according to which publications get the most exposure. The original study collected 382,139 data points from 186 news publications. However, this study chooses to focus on the United States, as it allows for a more controlled analysis within a single national context and because of the country's prominent role in shaping global discourse around the #MeToo movement and the COVID-19 pandemic. Once filtered, yields a dataset of 136,996 observations from 60 news US-based publications, ranging from 01/05/2005 to 06/06/2021.

Table 1 gives an overview of the variables of importance that are utilised in the analysis of this study. With the exception of the bias and polarity related variables, all data is sourced directly from the original Kaggle dataset. Explanations of how bias and polarity are quantified are provided in the subsequent Methodology section. It should be noted that this study defines bias not as a comparison between genders, but as the use of non-neutral, gender-marked, or stereotypical language in headlines about women. Drawing on feminist linguistics and computational media studies, it treats the frequency of such language, relative to neutral alternatives, as an indicator of framing bias.

Table 1: Variables in dataset used for analysis

| **Variable** | **Description** |
|---|---|
| Headline | The text of the news headline as published. |
| Headline, pre-processed | The text of the news headline as published after it has undergone pre-processing steps. |
| Date of Publication | The date on which the headline was published. |
| Publication Source | The name of the news outlet that published the headline. |
| Polarity Score | A value between -1 and 1 indicating how negative or positive a headline is, respectively. The score is calculated in this report from probability distribution of values derived from a pre-trained RoBERTa model. |
| Bias Word Count | A numeric score reflecting the degree of stereotypical language present in the headline. This score is calculated as the count of identified bias terms in the headline, as defined in the custom word dictionary. |
| Bias Proportion | A value between 0 and 1 indicating the proportion of biased language in a headline. This is calculated by dividing the bias word count by the number of words (excluding stopwords) in the headline. |

*Notes:* Polarity scores are derived from a three-class RoBERTa model. The custom dictionary used to identify gendered or stereotypical language is provided in Appendix B, Table B2.

Additionally, the *Women in Headlines: Bias* dataset includes a dictionary of English words classified as either *female words*, which indicate that the subject of a headline is female, or *female biased words*, which convey stereotypes or gender bias. This dictionary served as the foundation for the custom word list curated for this study. A detailed explanation of how the dictionary was adapted and applied can be found in Section 4, whilst examples of headlines with their respective bias and polarity scores can be found in Table 2.

### *3.1 Data Pre-processing*

An essential component of NLP is ensuring that textual data is appropriately prepared for analysis. In this dataset, the original form of each headline is retained, while an additional column containing the pre-processed version is included to support more robust semantic analysis. The pre-processing steps include: converting all text to lowercase, ensuring that words such as "Women" and "women" are treated equivalently; removing stopwords, which are common words such as "the" or "and" that are typically considered semantically uninformative; and applying lemmatisation to reduce words to their base or dictionary form (e.g., "worked" and "works" become "work"). These pre-processing tasks were implemented using standard Python libraries for NLP, including *spaCy* for lemmatisation and *NLTK* for stopword removal.

Care was taken to preserve the semantic integrity of the headlines throughout this process. Importantly, headlines that did not explicitly mention women or that lacked identifiable gender references were removed during data cleaning to ensure the final dataset remains focused and relevant to the research objectives.

### *3.2 Outcome Variables*

This study focuses on two primary outcome variables: female gender bias and sentiment polarity. The gender bias variable captures the presence of gendered or stereotypical language in headlines about women, based on a custom dictionary adapted from prior work on linguistic framing. The sentiment polarity variable measures the emotional tone of a headline, whether it is framed positively, negatively, or neutrally, using a transformer-based classifier. Together, these variables allow for a quantifiable, two-dimensional analysis of how women are portrayed in the news across time and media sources. Table 2 presents illustrative examples of headlines alongside their pre-processed forms, detected biased words, computed bias proportions, and polarity scores. These examples highlight how different linguistic elements contribute to variation across both dimensions. Full details on the construction of these variables are provided in the subsequent Methodology section.

Table 2: Example headlines and their respective outcome variables

| **Headline** | **Headline, preprocessed** | **Biased words detected** | **Bias proportion*** | **Polarity score** |
|---|---|---|---|---|
| Can a Mother's Pregnancy Diet Influence Her Child's Future Weight? | mother pregnancy diet influence child future weight | `pregnancy, diet, child, weight` | 0.5714 | -0.0670 |
| Jenna Fischer and Husband Welcome a Baby Girl—See Her Adorable Photo! | jenna fischer husband welcome baby adorable photo | `baby, adorable` | 0.2857 | 0.9598 |
| Five Year Old Girl Dead After Dad Threw Her Off a Bridge: Cops | five year old girl dead dad throw bridge cop | *n/a* | 0.0000 | -0.8976 |

*Notes:* *Bias proportion is the number of biased words detected over the number of words in the preprocessed headline.

Figure 1 presents the time series of average daily bias proportion and sentiment polarity in headlines from 2005 to 2021. The bias proportion, shown in the top panel, remains generally low, with the series reverting to a mean at approximately 0.05, but exhibits frequent spikes, particularly after 2017, suggesting sporadic surges in the use of gendered or stereotypical language. The sentiment polarity, displayed in the bottom panel, fluctuates around a slightly negative mean (approximately –0.1), indicating a generally negative tone in media portrayals of women. This trend is consistent with broader psychological evidence that negative information tends to attract more attention and have greater impact than positive content (Baumeister et al., 2001). Both series exhibit heteroscedasticity, i.e. the variance is not constant, with marked periods of heightened volatility interspersed with more stable intervals. A distinct period of lower volatility is observable from around 2014 to late 2018, followed by a notable increase in volatility across both dimensions. These shifts may correspond to external sociopolitical events and merit further investigation via structural break analysis or contextual annotation.

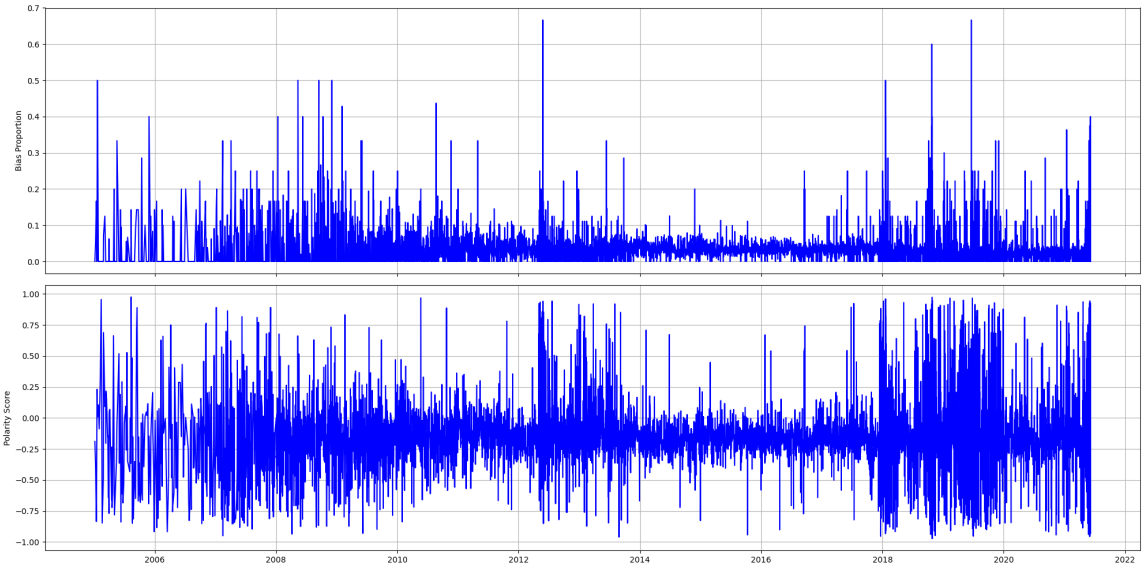



Figure 1: The evolution of bias (top panel) and polarity (bottom panel) measures from 2005 to 2021.

# 4 METHODOLOGY

## *4.1 Textual Analysis*

### *4.1.1 Defining Bias in this Study*

It is important to clarify what is meant by bias in the context of this thesis. In the broader literature, gender bias often refers to differential treatment or representation between men and women. However, this study focuses exclusively on headlines about women and does not compare them to male-focused headlines. As such, the term bias here does not denote contrastive bias between genders, but rather refers to the presence of non-neutral, gender-marked, or stereotypical language when discussing women. In this framework, a headline is considered biased if it contains terms that reinforce gender stereotypes, highlight gender unnecessarily, or invoke emotive or role-based associations that go beyond factual description. This approach aligns with work in feminist linguistics and computational media studies that treat the frequency of slanted or stereotyped language, relative to neutral alternatives, as an indicator of framing bias (Mills, 2008; Frenda et al., 2019). While this does not capture comparative bias, it allows us to measure whether language about women is becoming more or less slanted over time, offering insight into changing media portrayals.

### *4.1.2 Custom Dictionary of Gender Biased Vocabulary*

To identify gender bias in headlines, a dictionary-based approach is adopted, prioritising feasibility, transparency, and replicability. While contextual models such as BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019) offer advanced capabilities for detecting subtle biases, their application would require access to annotated corpora, fine-tuning infrastructure, and significant computational resources, beyond the scope of this thesis. Furthermore, prior research has shown that such models may inherit and reproduce societal biases embedded in their training data (Bolukbasi et al., 2016; Sheng et al., 2019).

Accordingly, this study utilises a curated dictionary derived from the *Women in Headlines* dataset hosted on Kaggle, which lists gendered and stereotypical terms commonly appearing in news coverage about women. However, upon review, the original dictionary was found to under-identify certain contextually gendered terms. For example, occupational titles such as *policewoman*, *stewardess*, and *actress* were not flagged as biased,

despite research in feminist linguistics showing that such terms mark gender unnecessarily and reinforce traditional role expectations in professional contexts (Mills, 2008, Chapter 3). Similar critiques of gender-marked language have also been noted in computational linguistic studies of media framing (Frenda et al., 2019).

To address this, the dictionary was revised to more clearly distinguish between purely referential gendered terms and contextually biased gender markers. A list of neutral female terms was first defined, including common pronouns and kinship terms such as *woman*, *women*, *she*, *her*, *daughter*, *mother*, and *wife*. These terms are frequently used in reference to female subjects without necessarily invoking stereotypes or role-based associations. This list is formalised in the *neutral_female_terms* category. Any word not in this list, particularly compound nouns or role descriptors that highlight gender when a gender-neutral alternative exists, was reclassified as part of the *female_biased_terms* column. The resulting list comprises explicitly gendered, stereotypically feminine, and often derogatory terms that reflect societal biases in how women are described, including references to appearance, emotion, domestic roles, and sexuality.

This reclassification acknowledges that while terms like *policewoman* may appear descriptive, they implicitly define the female referent as a deviation from a presumed male norm (*policeman* or more neutrally, *police officer*) (Fast and Horvitz, 2016). Similarly, *actress* and *stewardess* are marked counterparts to their unmarked male or gender-neutral versions (*actor*, *flight attendant*), and thus carry latent gender bias. By applying this updated classification scheme, the methodology aligns more closely with feminist linguistic critiques and enables a more rigorous detection of how gender is framed in news headlines. A full list of neutral and biased terms is included in Appendix B, Table B2.

While dictionary-based methods cannot fully account for the context in which words appear, they offer a transparent and interpretable approach to systematic language analysis (Garg et al., 2018; Jha and Mamidi, 2017). This methodology provides a solid empirical basis for tracking how gendered and stereotypical terms are used in headlines and aligns with the thesis's aim to investigate portrayals of women in the media. One limitation is that the dictionary does not include male-biased counterparts, which restricts its ability to capture contrastive framing or evaluate bias in contexts where gender is implied rather than explicitly stated. However, this study focuses exclusively on changes in the language surrounding women over time.

*4.1.3 Quantifying Gender Bias*

Once the dictionary of female gendered and biased terms was finalised, it was used to systematically quantify gender bias at the headline level. For each headline in the dataset, the number of biased words was counted by cross-referencing the tokenised text with the predefined list of biased terms from the custom dictionary. The distribution of raw biased word counts across headlines is summarised in Table 3. The majority of headlines (over 100,000) contain no biased terms, while a small subset contains multiple biased words, highlighting skewed patterns in how women are portrayed in certain contexts.

Table 3: Distribution of Headlines by Number of Biased Words

| **Number of Biased Words** | **Number of Headlines** |
|---|---|
| 0 | 108,771 |
| 1 | 24,657 |
| 2 | 4,638 |
| 3 | 658 |
| 4 | 66 |
| 5 | 7 |

While the raw count of biased terms could be used as a basic indicator, it does not account for headline length. Longer headlines naturally contain more words and therefore may have a higher chance of including biased terms simply by chance. To ensure comparability across headlines of differing lengths, the raw biased word count is normalised by the total number of non-stopword tokens in each headline. This method ensures that the resulting metric reflects the density of biased language relative to the meaningful content of the headline, rather than its absolute presence (Garg et al., 2018). For example, a headline containing two biased terms in a five-word headline represents a much stronger signal than two biased terms in a fifteen-word headline, capturing the density of biased language relative to the meaningful content of the headline.

The normalisation is based on the preprocessed version of each headline, specifically, the set of tokens remaining after removing stopwords and punctuation. This choice ensures that the denominator reflects only the semantically meaningful content, rather than filler words that contribute little to the overall message or framing. For example, arti-

cles, conjunctions, and prepositions such as "the," "and," or "in" are common across all headlines but do not carry interpretive weight. Including them would dilute the bias proportion and obscure the relationship between biased terms and the substantive linguistic content. Preprocessing therefore helps isolate the proportion of bias within the words most likely to shape reader perception, aligning with standard practices in lexical bias detection.

However, it should be acknowledged that not all additional words are irrelevant, and many simply provide factual context (e.g., location, time) rather than mitigating the interpretive weight of the biased term. For example, consider the short headline "Policewoman injured in crash" and a longer variant "Policewoman injured in crash in Lund yesterday". Both contain the same biased term ("policewoman"), but the former yields a higher bias proportion due to its brevity, indicating a denser concentration of gendered language. As such, while the bias proportion helps standardise measurement across headlines, it does not offer a perfect proxy for perceived bias. Alternative approaches, such as using raw counts or applying position-weighted metrics that emphasise early-occurring terms, could offer complementary perspectives. Nonetheless, for this study, I prioritise a metric that is both interpretable and comparable, and that aligns with prior work in lexical bias detection[2].

Figure 2 visualises the frequency of biased words across the dataset. The most frequently occurring terms include *lady*, *baby*, and *mom*, suggesting a persistent emphasis on familial or infantilising frames. Words like *pregnant*, *sex*, and *body* also appear prominently, reflecting common patterns in media framing that associate women with reproductive roles, physical appearance, or sexualisation. The long-tail distribution shows that while many biased terms occur infrequently, a small set of stereotyped language dominates media discourse.

[2]Other methods for constraining the bias count, such as min-max scaling or division by the maximum number of biased terms observed, were not used, as they do not account for the linguistic context in which the terms appear. Min-max scaling, for instance, would map all values to a [0,1] range based on dataset-wide extrema, but would not reflect whether a single biased term in a very short headline constitutes a higher proportion of bias than multiple biased terms in a longer headline. Similarly, standardising or rank-transforming the counts would obscure the intuitive interpretability of bias as a per-word proportion, which is particularly important when analysing journalistic text, where conciseness is a stylistic norm (Frenda et al., 2019).

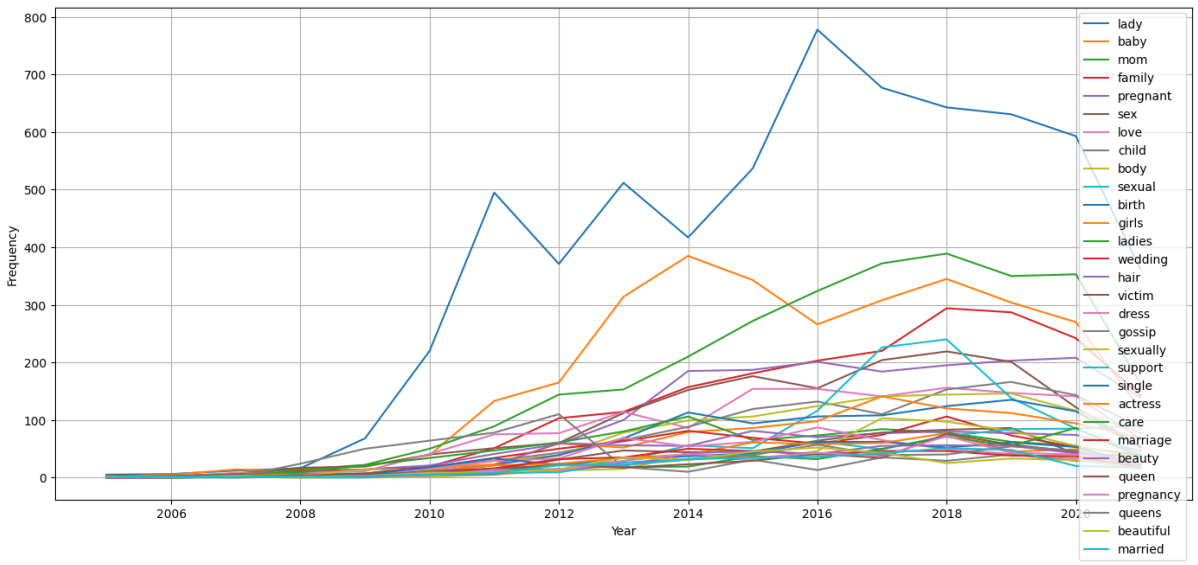



Figure 2: The 30 most frequently occurring female gender-biased words across all headlines over time.

### *4.1.4 Measuring Polarity*

To quantify the sentiment of each headline, this study employs the *cardiffnlp/twitter-roberta-base-sentiment* (CardiffNLP) model, a RoBERTa-based transformer pre-trained on Twitter and fine-tuned for sentiment classification (Barbieri et al., 2022). This model was selected due to its strong performance on short-form texts and its three-way classification scheme (negative, neutral, and positive) which enables more nuanced sentiment analysis compared to binary alternatives such as SiEBERT.

Despite being trained on Twitter data, where language is generally informal and highly contextual, the CardiffNLP model performs well on other domains with short, syntactically compressed texts. Headlines, while more formal, share similar linguistic constraints, making this model a suitable fit. Unlike lexicon-based methods like VADER, which often struggle with negation, sarcasm, or unusual syntax, transformer-based models such as RoBERTa leverage deep contextual embeddings to yield more accurate sentiment interpretation (Purohit et al., 2025; Mohammad, 2020; Balahur, 2013).

The model is applied directly to the unprocessed textual headline data, as it is designed to interpret raw language inputs without requiring extensive pre-processing. RoBERTa outputs a probability distribution over three sentiment classes: negative ($-1$), neutral ($0$), and positive ($+1$). To derive a continuous sentiment score, the expected value of this distribution is computed as follows:

$$\text{Polarity Score} = Pr(\text{Positive}) \cdot 1 + Pr(\text{Neutral}) \cdot 0 + Pr(\text{Negative}) \cdot (-1) \quad (1)$$

This results in a score ranging from $-1$ (strongly negative) to $+1$ (strongly positive), with 0 representing a neutral sentiment. For instance, the headline "Woman dies after tummy tuck at Miami surgery center" receives a polarity score of $-0.93$, reflecting its strongly negative tone. In contrast, "Kaley Cuoco Emotionally Praises Person Who Returned Her Lost Wallet: 'There Is Good in the World'" scores $+0.89$, indicating highly positive sentiment. Headlines that convey factual events without emotional emphasis, such as "Photo of firefighter cradling little girl after car accident goes viral", tend to receive scores near zero, in this case, $+0.09$.

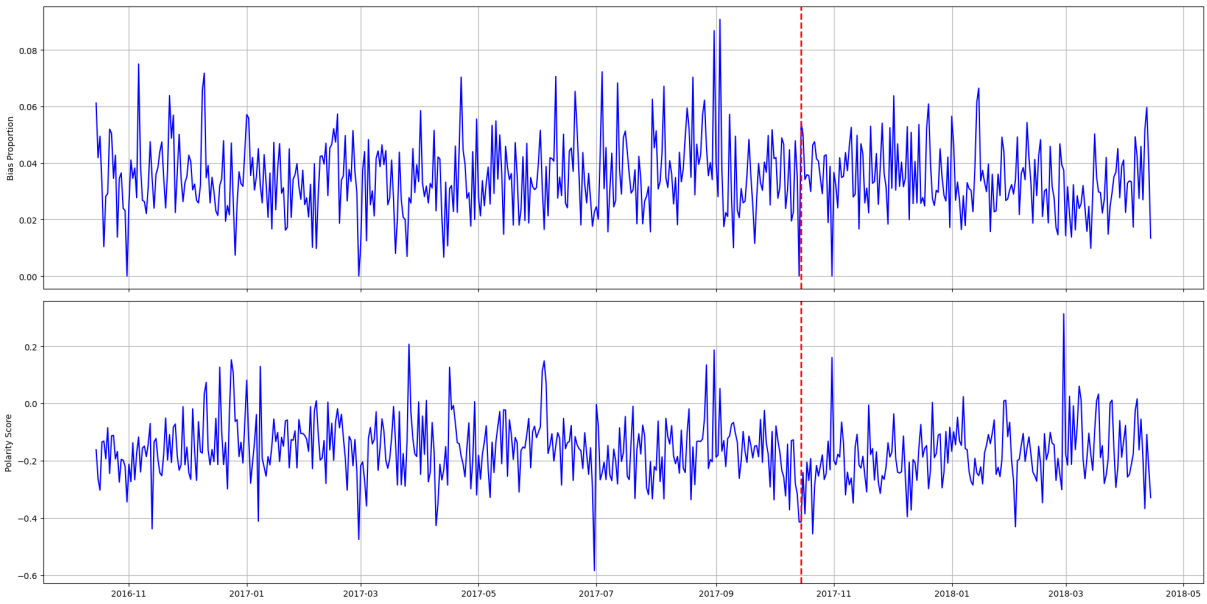
The inclusion of a neutral class allows for better detection of balanced or factual language, a key feature in journalistic writing. Future work could explore fine-tuning transformer models directly on headline corpora to further tailor sentiment detection to the genre.

### *4.2 Time Series Analysis*

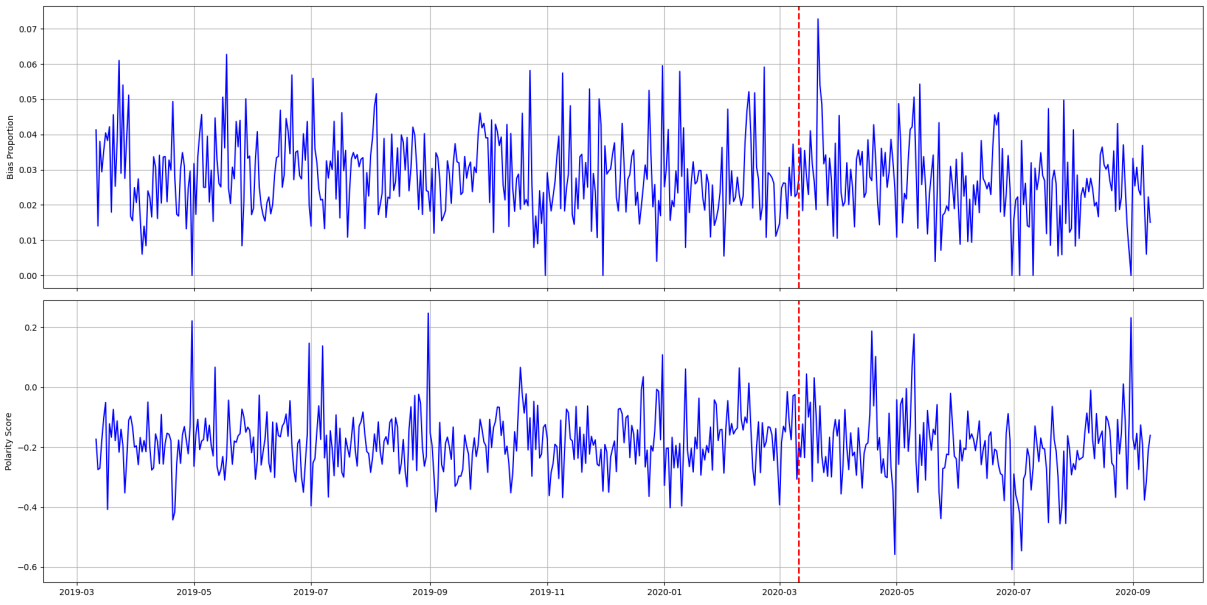
This study adopts a hybrid time series analysis approach to investigate whether the sentiment polarity and biased language used in news headlines about women shifts in response to major societal events. The methodology combines data-driven changepoint detection with event-based pre/post analysis, to both identify structural breaks in the data and assess the potential causal influence of events such as the #MeToo movement and the COVID-19 pandemic.

Figure 3 presents the time series of bias proportion (top) and sentiment polarity (bottom) over a 12-month period preceding and a 6-month period following two major societal events: the onset of the #MeToo movement (panel a) and the declaration of the COVID-19 pandemic (panel b). A red vertical line marks the date of each event.

Overall, the series exhibit mean-reverting behaviour, with bias proportion tending toward a low positive mean and sentiment polarity fluctuating around a mildly negative mean. However, visual inspection reveals fluctuations in variance, such as localised peaks and dips in volatility around both event dates. These may correspond to transient changes in media framing or tone. Notably, there is a peak in bias in the lead-up to #MeToo,




(a) #MeToo bias and polarity measures




(b) COVID-19 bias and polarity measures

Figure 3: The bias proportion (top) and sentiment polarity (bottom) measures the 12 months leading up to (a) #MeToo (b) and COVID-19, indicated by the red line, and the 6 months that followed.

followed by temporary dips, while polarity also rises before the event, gradually declines, and then peaks again shortly after its onset. These fluctuations may reflect shifts in media focus, heightened attention to gender-related issues, or changing editorial priorities, although such interpretations remain speculative based on visual trends alone.

In contrast, COVID-19 is associated with an immediate and pronounced spike in bias, whereas polarity remains initially stable before gradually declining—potentially indicating a shift toward more negative sentiment as the broader societal impact of the pandemic became more apparent. Again, while these patterns are suggestive, their interpretation must be approached with caution in the absence of formal statistical confirmation.

While these patterns suggest potential structural changes, their statistical significance cannot be determined through visual inspection alone. Formal changepoint detection is therefore applied in the following sections to evaluate whether these apparent shifts represent meaningful breaks in the underlying data-generating process.

The correlation between the bias and polarity series is found to be negligible at 0.03, however both can be studied to give a multi-dimensional analysis of how women are portrayed in the media. To preserve the original structure and volatility of the data, no smoothing or moving average was applied prior to changepoint analysis. The raw daily time series for both bias proportion and sentiment polarity are used directly in the analysis to ensure that abrupt shifts, especially those related to sudden media responses to events, are not weakened. This approach maintains the fidelity of high-frequency fluctuations, which are relevant for identifying real-world changepoints. The methods describe in this section will be applied to both bias and polarity measures.

### *4.2.1 Unsupervised Detection of Structural Breaks*

To identify structural breaks in the time series without relying on prior assumptions, this study employs the *ruptures* Python package. The package enables the identification of points in time where the statistical properties—such as the mean or variance—change significantly. The *PELT* (Pruned Exact Linear Time) algorithm is selected due to its computational efficiency and its ability to detect multiple changepoints without needing to pre-specify their number. PELT works by minimising a cost function within each segment, applying a penalty term to avoid overfitting via excessive segmentation. Its pruning strategy ensures linear scalability, making it particularly suited to high-frequency

data such as daily news headlines (Killick et al., 2012). Two key parameters must be specified: the cost function, which defines the type of structural changes to detect, and the penalty term, which balances model fit against complexity. Following Truong et al. (2020), this study implements both the L2 and RBF cost functions to capture different types of changes, shifts in the mean as well as shifts in both mean and variance.

**L2 Cost Function (Mean Shift Detection):**

The L2 cost function is designed to identify changes in the mean of a time series while assuming constant variance. This is particularly appropriate when tracking shifts in average bias or polarity in headlines, where level changes may reflect long-term shifts in media tone. For instance, a sustained increase in gendered language following a major societal event may manifest as a mean shift. The L2 function is also advantageous for its interpretability, facilitating pre- and post-event comparisons aligned with the study's aims.

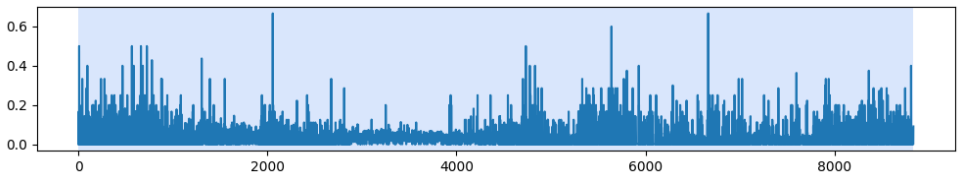
**RBF Cost Function (Mean and Variance Shift Detection):**

In contrast, the RBF (Radial Basis Function) cost function captures shifts in both mean and variance. This added sensitivity is valuable when analysing volatile media dynamics, such as surges in emotional language during the MeToo movement or the COVID-19 pandemic. By accounting for variance, RBF provides a more nuanced view of structural change, capable of detecting regime shifts in both tone and volatility.

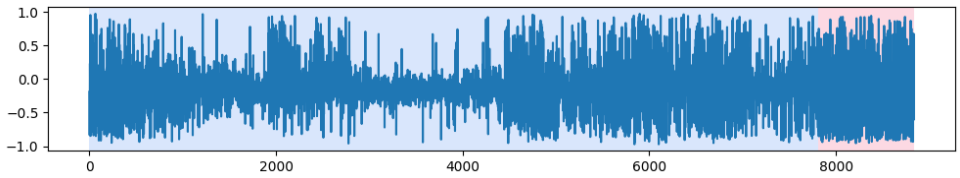
**Penalty Term and Robustness:**

The penalty term is calibrated through a combination of visual inspection and Bayesian Information Criterion (BIC) minimisation, which penalises model complexity more strongly than the Akaike Information Criterion (AIC). This conservative approach helps reduce false positives, especially under the null hypothesis of no structural change. While BIC is prioritised, AIC is employed in robustness checks to assess the sensitivity of changepoint detection to penalty strength. As expected, lower penalty values increase sensitivity to small fluctuations, while higher values reduce over-segmentation. Based on inspection of BIC and AIC behaviour (see Appendix C1), optimal penalty values of 2 for bias and 3 for polarity were selected.

As anticipated, the RBF model identifies more breakpoints than the L2 model (see Figure 4 and Figure 5). For example, the L2 segmentation identifies a single breakpoint in polarity (20 May 2021), while the RBF model detects multiple shifts across both bias

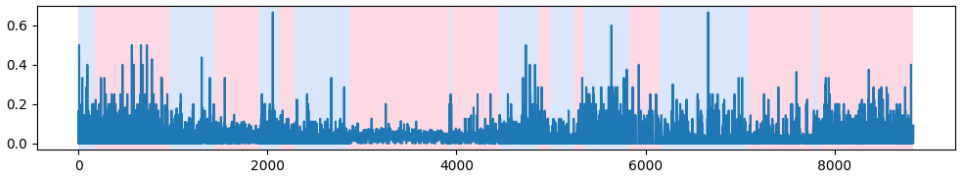



(a) Bias proportion

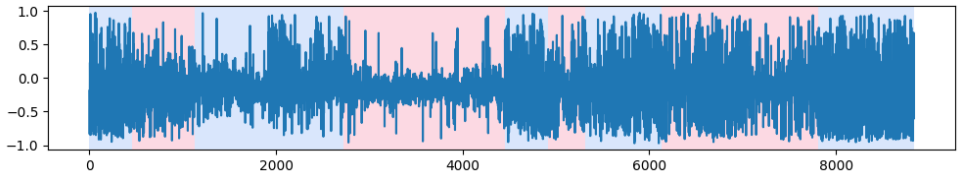



(b) Sentiment polarity

Figure 4: Mean breakpoints in the (a) bias and (b) polarity measures identified by unsupervised PELT algorithm with a L2 cost function.




(a) Bias proportion




(b) Sentiment polarity

Figure 5: Mean and/or variance breakpoints in the (a) bias and (b) polarity measures identified by unsupervised PELT algorithm with a RBF cost function.

and polarity. A full list of identified changepoints is presented in Table 5 in the Results section.

A key consideration in unsupervised changepoint detection is the risk of overfitting, mistaking noise for structure. A well-calibrated penalty term should yield few or no breakpoints in a stationary series. Nonetheless, even with BIC tuning, the more sensitive RBF function may detect changes driven by variance alone. This highlights the importance of interpreting detected changepoints in light of external events or broader trends.

While the penalty term helps control model complexity, it does not formally correct for multiple testing. As the algorithm evaluates many potential changepoints, the risk of false positives increases with longer or higher-frequency series. PELT does not adjust $p$-values or error rates across comparisons, so some breakpoints may arise by chance. This underscores the need for cautious interpretation, particularly when breakpoints lack external validation or cross-method agreement. Future work could incorporate corrections, such as permutation tests or bootstrapped null models.

To enhance interpretive confidence, both cost functions are applied to the same series. Changepoints detected by both L2 and RBF suggest more robust structural shifts. In contrast, RBF-only breakpoints may reflect variance changes that, while potentially meaningful, are less easily contextualised.

#### *4.2.2 Event-Based Analysis*

While changepoint detection enables the unsupervised identification of structural shifts in the bias and sentiment polarity series, it does not explain why these shifts occur. The method is exploratory in nature and does not attribute changes to specific external events. To further investigate whether meaningful changes in media tone align with real-world developments, this approach is complemented with event-based analysis. Specifically, I examine pre- and post-event dynamics around two major historical moments that are expected to influence media portrayals of women: the onset of the #MeToo movement on 15th October 2017, and the declaration of the COVID-19 pandemic by the World Health Organization on 11th March 2020. For each event, a pre-event window of 12 months is defined to ensure sufficient data to estimate underlying trends while avoiding confounding from long-term drift. A post-event window of up to 12 months used across horizons of 1, 3, 6 and 12 months. By focusing on known events, this approach allows for a more

targeted assessment of how media framing may have shifted in response to broader social and political contexts, while avoiding any unwarranted causal claims.

**Check for Stationarity**

Before modelling, it is essential to assess the stationarity of the time series. Many time series models assume stationarity in mean and variance. Non-stationary series can lead to unreliable forecasts and spurious conclusions in mean or variance tests. To test for stationarity, the Augmented Dickey-Fuller (ADF) test (Mushtaq, 2011) is applied to both the bias and sentiment polarity series in their pre-event windows.

Stationarity was assessed in the 12-month windows preceding both the #MeToo and COVID-19 events using the ADF test. For the #MeToo window, the bias proportion series had an ADF statistic of –8.77 ($p < 0.001$), and the polarity series –19.85 ($p < 0.001$). In the COVID-19 window, the bias and polarity series produced ADF statistics of –14.29 and –26.68, respectively (both $p < 0.001$). In all cases, the test statistics were well below the 1% critical values, providing strong evidence of stationarity. As a result, differencing is not required in order to fit an AutoRegressive Moving Average (ARMA) model. Given the absence of long-term trends or seasonality, the next step is to proceed with forecasting using ARMA models with constant mean terms to evaluate potential structural changes in the post-event periods.

**Fit Model to Pre-Event Data**

This study chooses to begin by fitting an ARMA model as such models as considered to be amongst the strongest time series modelling techniques (Hyndman and Athanasopoulos, 2018, Chapter 8). They combine auto-regressions with a moving average over error terms, capturing both the persistence in the series and the structure in the forecast errors. An ARMA model is defined by two components: an AutoRegressive (AR) model of order $p$ and a Moving Average (MA) model of order $q$.

An AR($p$) model uses the past $p$ values in the time series to apply a regression, formulated as:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t \tag{2}$$

where $c$ is the intercept (or constant term), and $\epsilon_t$ represents the error term at time $t$.

An MA($q$) model, in contrast, regresses the forecast errors, and is expressed as:

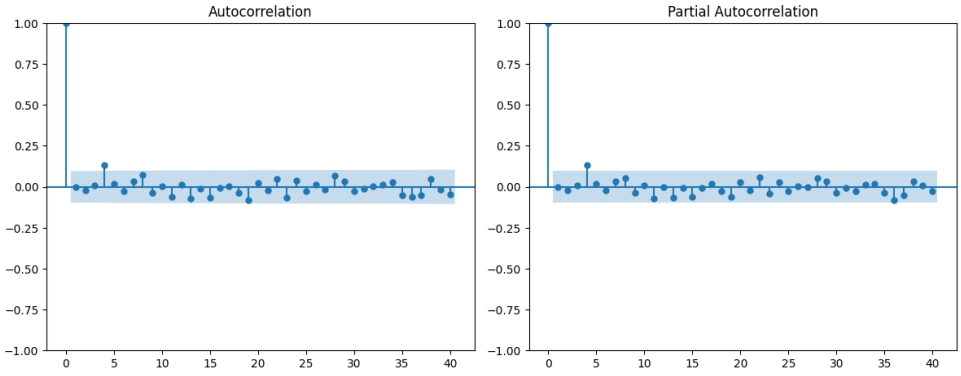
$$y_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} \tag{3}$$

The ARMA model combines both AR and MA components, enabling the modelling of both autocorrelation in the series and structure in the residuals. It is particularly suitable for stationary time series. Given that stationarity of the series has been established, differencing is not required, and the ARMA model may be applied directly without extending to an ARIMA specification.

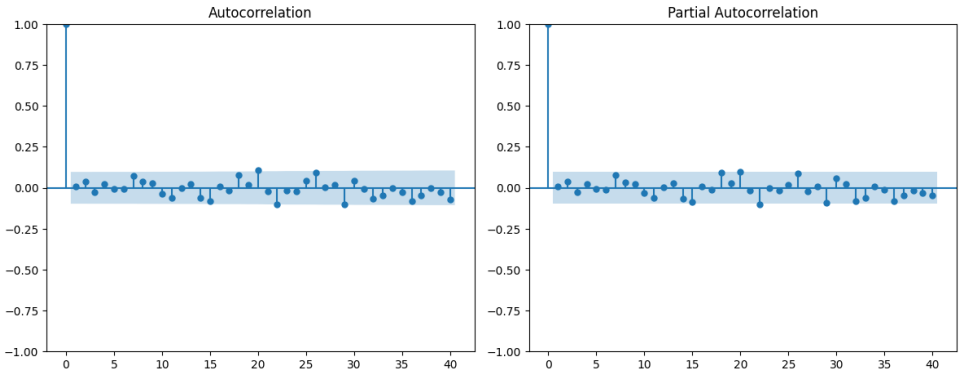
To identify appropriate ARMA model structures for the pre-event time series, a multi-step diagnostic approach is employed. The autocorrelation function (ACF) and partial autocorrelation function (PACF) plots are first examined to explore serial dependence in the raw data (Burnham and Anderson, 2002). The ACF and PACF plots of the bias and polarity series in Figure 6 suggest that both series are weakly autocorrelated. The ACF shows a single large spike at lag 1, followed by small, alternating positive and negative values that quickly decay. This mild oscillating pattern can indicate a short-memory process, potentially arising from an ARMA(1,0) structure with a negative coefficient. The PACF also displays a sharp cutoff after lag 1, with the remaining lags fluctuating around zero within the 95% confidence bands. This pattern aligns with the behaviour of a stationary autoregressive process of order one, ARMA(1,0), and suggests no additional higher-order lags are needed.

Model adequacy is assessed by examining the residuals and their squared values. ACF and PACF plots of residuals are used to check for any remaining autocorrelation, while the Ljung-Box test formally evaluates whether the residuals approximate white noise. A non-significant Ljung-Box statistic indicates that the chosen model has successfully captured the temporal structure in the data (Ljung and Box, 1978). The squared residuals are assessed for heteroscedascity and when volatility clustering is detected, Generalised Autoregressive Conditional Heteroscedasticity (GARCH) models are employed to capture time-varying variance. These models allow the analysis to assess changes not only in central tendency but also in the stability and dispersion of polarity and bias over time. The ARCH tests performed on the squared residuals for both series and both events returned p-values well above the 0.05 threshold, indicating that volatility is not time-dependent in these series and a GARCH specification is unnecessary.

As the core aim of this event analysis is to evaluate whether structural shifts occur following major events, the fitted pre-event ARMA and/or GARCH models are used to forecast the post-event period under the assumption of no change. Forecasts are then



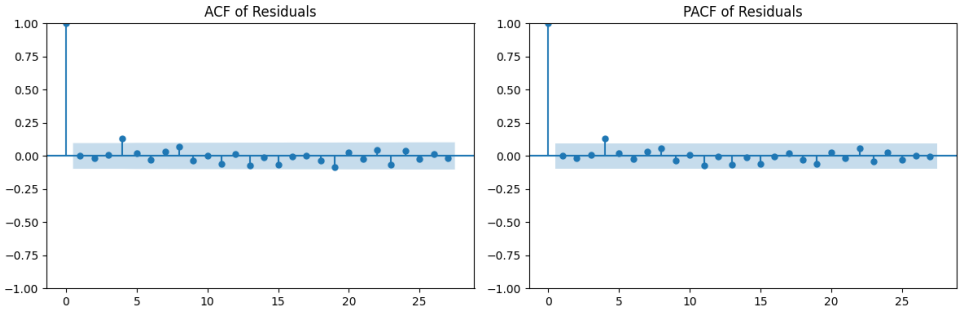

(a) Bias proportion



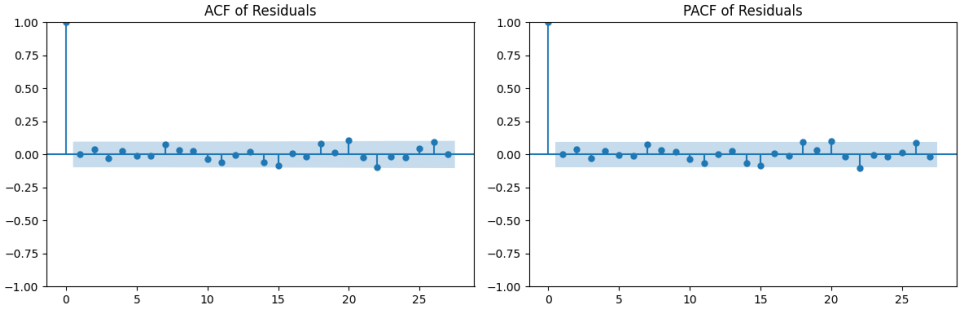

(b) Sentiment polarity

Figure 6: ACF and PACF plots for raw bias and polarity series in the pre-#MeToo period

compared to actual post-event values to test for significant deviations in both the mean and the variance of the series.




(a) Bias proportion




(b) Sentiment polarity

Figure 7: ACF and PACF plots for residuals of ARMA(1,0) models fitted on (a) bias and (b) polarity series in the pre-#MeToo period

In the pre-#MeToo window, the ACF and PACF plots for the bias proportion series show a sharp spike at lag 1 followed by values within the 95% confidence bounds, consistent with a weak autoregressive structure. An ARMA(1,0) model was therefore fitted, incorporating a constant mean term and a single AR component. However, the estimated AR(1) coefficient was not statistically significant ($\beta = -0.0034, p = 0.951$), indicating that the series behaves largely as a mean process. Nonetheless, the inclusion of the AR term, along with the intercept, enables the model to serve a meaningful forecasting role under the null hypothesis of no change. Residual diagnostics provided strong support for model adequacy: the ACF and PACF of the residuals showed no evidence of autocorrelation (Figure 7a), and the Ljung–Box test at lag 10 returned a non-significant result ($p = 0.354$), confirming that residuals approximate white noise. While the AR(1) term may not be significant on its own, it is retained to maintain consistency with the model structure across series and to avoid underfitting. The model captures the low average level of bias in the pre-event period (intercept = 0.0358, $p<0.001$), and no substantial

misspecification is detected.

For the sentiment polarity series, the ACF and PACF plots display an initial spike at lag 1 with no significant lags thereafter, again pointing toward a weak autoregressive structure. Accordingly, an ARMA(1,0) model was estimated, including a constant term and one AR lag. As with the bias series, the ARMA(1,0) coefficient was not statistically significant ($\beta = 0.0101$, $p = 0.860$), suggesting the model primarily captures a stationary mean. The ARMA structure appears adequate, with the ACF and PACF plots of the residuals showing no systematic autocorrelation (Figure 7b) and the Ljung–Box test at lag 10 yields a p-value of 0.902. The model's intercept term is –0.1444 ($p<0.001$), reflecting the generally negative polarity scores in the pre-event period. In line with the bias model, the ARMA(1,0) specification is retained for its simplicity, diagnostic validity, and consistency across the two series.

The ACF and PACF plots for the pre-COVID window (see Appendix C, Figure C2a) exhibit a similar autocorrelation structure to that observed in the pre-#MeToo period. Consequently, the same methodological approach of model identification using AIC and BIC, followed by ARMA model fitting and diagnostic evaluation, is applied to both events. An ARMA(1,0) model is also fitted to the pre-COVID-19 window. Although the AR(1) coefficients are not statistically significant, the inclusion of the autoregressive term along with the intercept enables the model to function effectively for forecasting under the null hypothesis of no structural change. This maintains methodological consistency and avoids potential underfitting.

**Forecasting and Evaluation of Structural Changes**

To evaluate whether structural changes occurred in the sentiment polarity and bias proportion series following key events, forecasts are generated using the models fitted to the 12-month pre-event windows. These models are employed to forecast post-event values under the null hypothesis of no change. Visual inspection of time series plots allows for intuitive assessment of divergence between expected and observed behaviour. To formally test for changes in distributional moments, the mean and variance of each series are calculated for both pre- and post-event windows. Differences in mean are evaluated using independent-sample $t$-tests, while $F$-tests are used to assess differences in variance. A statistically significant change in either the mean and/or the variance is interpreted as evidence of a structural shift in how women are represented in headlines following

the specified event. The forecasted values and statistical test results are presented and interpreted in the next section.

By combining changepoint detection with event-based analysis, this dual approach enables both the discovery of structural shifts and their validation against known historical events.

# 5 RESULTS

To investigate changes in gendered and stereotypical language in media headlines over time, this study evaluates whether significant structural changes in bias proportion and sentiment polarity occurred around two major social events: the #MeToo movement and the COVID-19 pandemic. These events were chosen for their significant cultural and societal impacts, which plausibly influenced media narratives concerning women. Specifically, the following null hypothesis is tested:

$H_0$: *The bias and polarity of gendered language in headlines about women have remained consistent over time.*

Rejection of this hypothesis in favour of $H_1$ would indicate that such language has changed in frequency and/or tone, especially the focal events for this study, #MeToo and COVID-19.

A significance test assesses whether an observed effect, such as a shift in mean or variance, is likely to have occurred by chance under a null hypothesis of no change. The resulting p-value quantifies the probability of observing a test statistic at least as extreme as the one computed, assuming the null hypothesis is true. In this study, $t$-test statistics are used to evaluate shifts in mean, while $F$-test statistics assess those in variance. A low p-value, typically below 0.05, indicates that the observed change is unlikely to be due to random variation alone and is therefore considered statistically significant.

Figure 8 and figure Figure 9 present the actual and forecasted values of bias proportion and sentiment polarity around, respectively, the #MeToo event boundary in October 2017 and that of COVID-19 in March 2020. Forecasts were generated using ARMA(1,0) models trained on a 12-month pre-event window. Table 4 reports the mean and variance of actual and forecasted series across 1-, 3-, 6-, and 12-month post-event horizons.

## *5.1 #MeToo Period Results*

The bias proportion during the #MeToo period did not show statistically significant deviations from forecasted values at any horizon (all $p$>0.23), therefore it is unlikely that these changes in bias are anything other than random. While slight differences in means and variances were observed, they were not sufficient to reject the null hypothesis of stability. This suggests that the overall frequency of biased language remained relatively

consistent, despite the increased media attention on women's experiences.

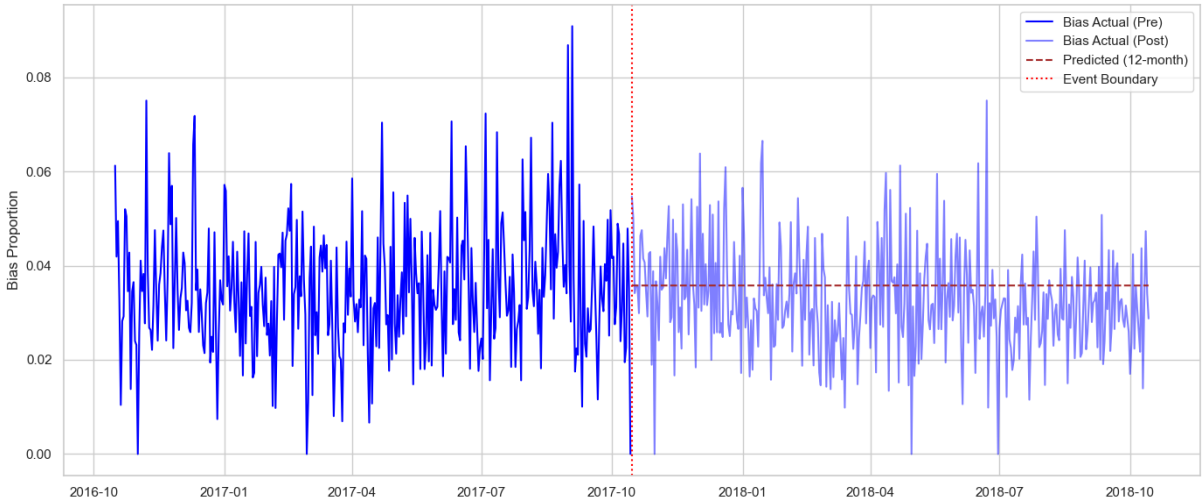
In contrast, sentiment polarity exhibited substantial and statistically significant shifts. Within one month post-event, the mean polarity became significantly more negative than predicted ($t = -2.91$, $p = 0.007$), and this effect persisted at the 3-month ($t = -3.56$, $p = 0.006$) and 12-month ($t = -2.01$, $p = 0.046$) horizons. This downward shift in polarity suggests that although women's stories were more prevalent in headlines, the tone in which they were portrayed became more emotionally charged, possibly reflecting a reactive media environment engaging with serious topics such as abuse, harassment, and trauma.

Variance in polarity also increased substantially, with actual values exceeding forecasted variance by several orders of magnitude. $F$-tests confirmed this with extremely large $F$-statistics across all horizons (e.g., $F = 692.24$, $p<0.00001$ at 1 month), indicating heightened emotional variability in coverage.

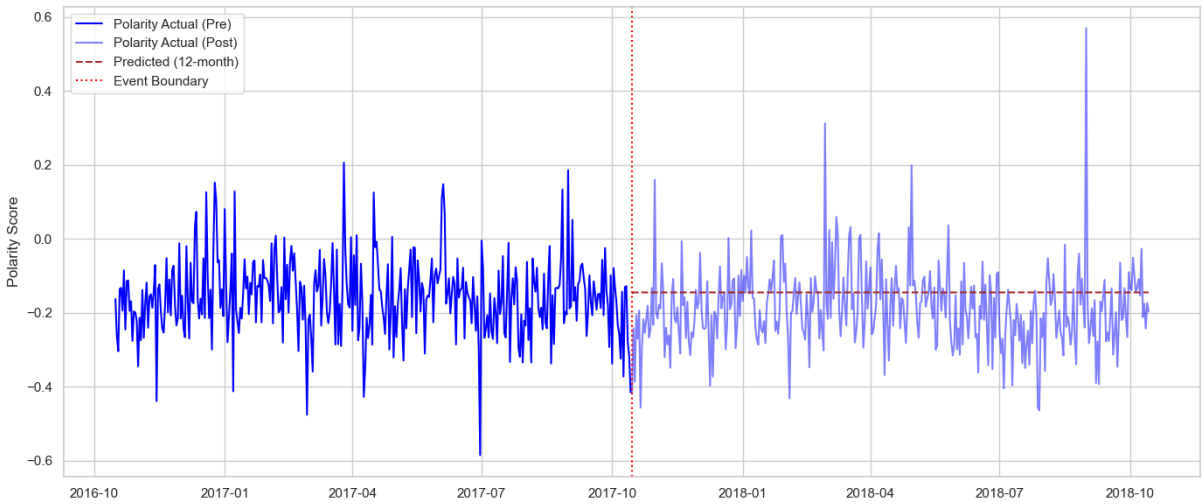
### *5.2 COVID-19 Period Results*

In this period, the bias proportion increased slightly compared to forecasted values for each horizon, however the difference in the mean biases only becomes statistically significant at the 12-month horizon ($t = -2.02$, $p = 0.045$). This result may reflect a re-entrenchment of traditional gender roles in the media during lockdowns and economic uncertainty, potentially reinforcing stereotypes about women in domestic and caregiving roles.

Polarity, while persistently negative across the post-COVID period, did not exhibit statistically significant mean shifts at any horizon (all $p > 0.11$). However, as with the #MeToo period, the variance of polarity was markedly greater than predicted, with F-statistics again confirming structural changes (e.g., $F = 12368.22$, $p < 0.000001$ at 1 month). This increased variability may point to a less uniform media narrative, encompassing a mix of both empowering and regressive stories about women during the pandemic.

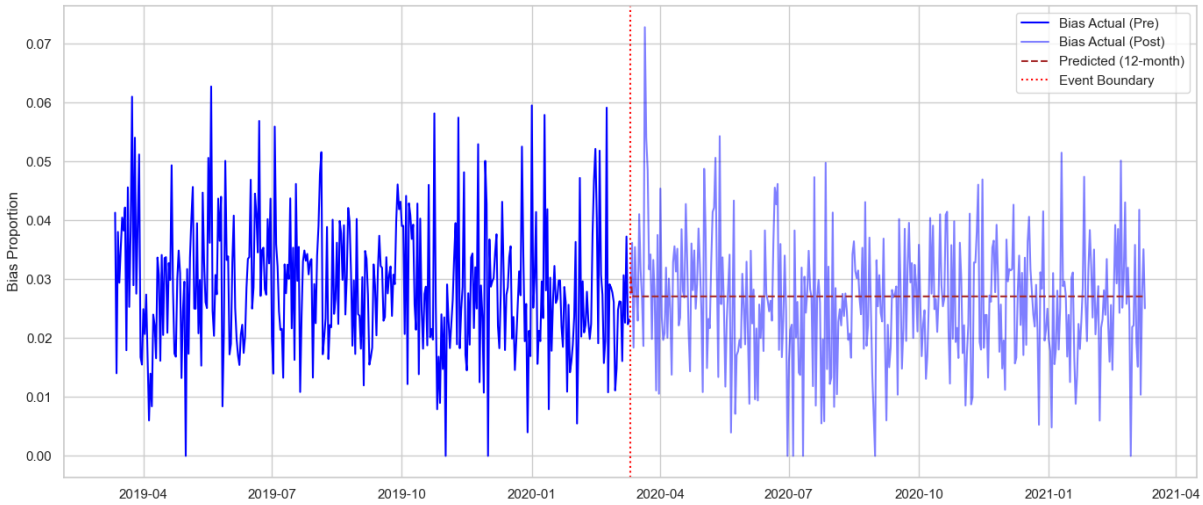



(a) Bias proportion

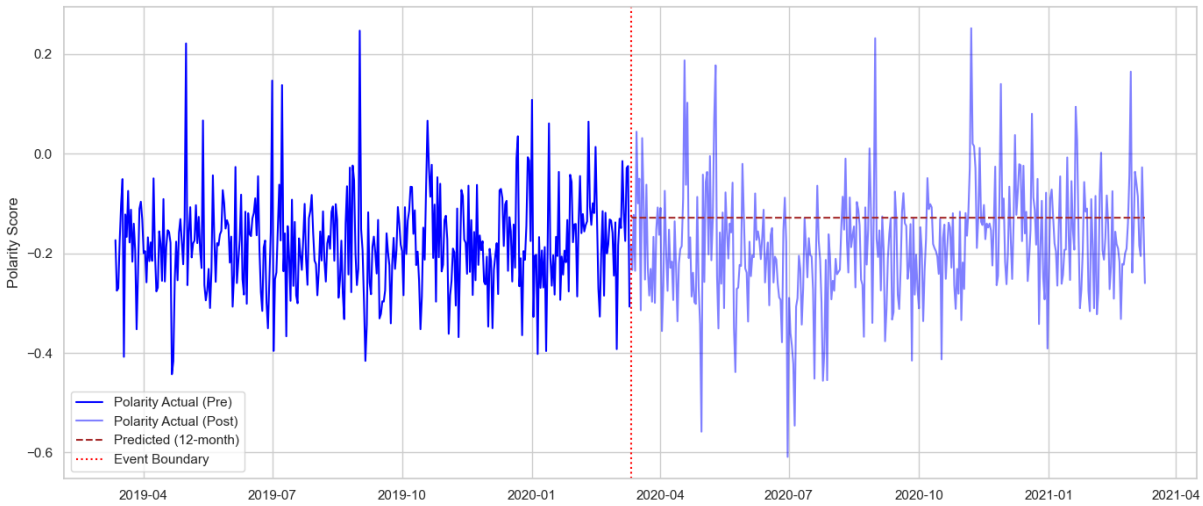



(b) Sentiment polarity

Figure 8: #MeToo period forecasted values




(a) Bias proportion




(b) Sentiment polarity

Figure 9: COVID-19 forecasted values

Table 4: Mean and variance of actual and forecasted values post-event

| Horizon | Mean Forecast | Mean Actual | Mean Diff. | $t$-stat (p-value) | Var. Forecast | Var. Actual | Var. Diff. | $F$-stat (p-value) |
|---|---|---|---|---|---|---|---|---|
| **#MeToo – Bias Proportion** | | | | | | | | |
| 1 month | 0.034666 | 0.037149 | -0.00248 | 1.225 (0.231) | <0.000001 | 0.000121 | 0.000120 | 1258.35 (<0.00001) |
| 3 months | 0.034707 | 0.035642 | -0.00094 | 0.765 (0.446) | <0.000001 | 0.000134 | 0.000134 | 4151.92 (<0.00001)) |
| 6 months | 0.034717 | 0.034155 | -0.00056 | -0.652 (0.515) | <0.000001 | 0.000133 | 0.000133 | 8274.53 (<0.00001)) |
| 12 months | 0.034723 | 0.032737 | -0.00064 | -0.746 (0.457) | <0.000001 | 0.000126 | 0.000125 | 8436.16 (<0.00001)) |
| **#MeToo – Sentiment Polarity** | | | | | | | | |
| 1 month | -0.165620 | -0.225162 | -0.059542 | -2.906 (0.007) | 0.000018 | 0.012767 | 0.012749 | 692.24 (<0.00001)) |
| 3 months | -0.165013 | -0.201151 | -0.036138 | -3.561 (0.006) | 0.000006 | 0.009380 | 0.009374 | 1513.88 (<0.00001)) |
| 6 months | -0.164862 | -0.179441 | -0.014579 | -1.863 (0.064) | 0.000003 | 0.011084 | 0.011081 | 3571.08 (<0.00001)) |
| 12 months | -0.164785 | -0.180436 | -0.015651 | -2.010 (0.046) | 0.000002 | 0.011095 | 0.011093 | 3614.50 (<0.00001)) |
| **COVID-19 – Bias Proportion** | | | | | | | | |
| 1 month | 0.027802 | 0.029928 | 0.00213 | 0.880 (0.386) | <0.000001 | 0.000175 | 0.000175 | 30917.53 (<0.000001) |
| 3 months | 0.027812 | 0.028253 | 0.00044 | 0.358 (0.721) | <0.000001 | 0.000137 | 0.000137 | 72411 30 (<0.000001) |
| 6 months | 0.027814 | 0.026334 | -0.00148 | -1.729 (0.0856) | <0.000001 | 0.000132 | 0.000132 | 13908.52 (0.00) |
| 12 months | 0.027815 | 0.026134 | -0.00171 | -2.016 (0.0452) | <0.000001 | 0.000133 | 0.000133 | 143178.66 (0.00) |
| **COVID-19 – Sentiment Polarity** | | | | | | | | |
| 1 month | -0.188679 | -0.179317 | 0.00936 | 0.512 (0.613) | <0.000001 | 0.010044 | 0.010044 | 12368.22 (<0.000001) |
| 3 months | -0.188810 | -0.181574 | 0.00724 | 0.554 (0.581) | <0.000001 | 0.015379 | 0.015378 | 56260.76 (<0.000001) |
| 6 months | -0.188843 | -0.201708 | -0.01286 | -1.412 (0.160) | <0.000001 | 0.014932 | 0.014932 | 105997.15 (0.00) |
| 12 months | -0.188860 | -0.182740 | -0.01424 | -1.585 (0.115) | <0.000001 | 0.014849 | 0.014849 | 110790.79 (0.00) |

*Notes:* Each row compares actual and forecasted values at a given horizon after the event. $t$-statistics and p-values come from paired sample $t$-tests, and variance differences are assessed using formal $F$-tests, as reported in the final column.

*5.3 Structural Breaks Detected Using ruptures*

To supplement the forecast comparison approach, change point detection was applied using the *ruptures* package with both L2 and RBF models to identify endogenous structural breaks in the full series (Table 5). The L2 model detected a single break in polarity in May 2021, while the more sensitive RBF model identified numerous breakpoints across both bias and polarity series. Crucially, several breakpoints cluster around key periods—late 2017 to early 2018 (coinciding with #MeToo) and early 2020 (onset of COVID-19), supporting the hypothesis that these events correspond with structural shifts in media discourse.

Table 5: Breakpoints in Bias and Polarity for L2 and RBF Models

| **Model** | **Breakpoints in Bias** | | | **Breakpoints in Polarity** | |
|---|---|---|---|---|---|
| L2 | n/a | | | 2021-05-20 | |
| RBF | 2006-08-26, | 2009-09-14, | 2011-01-10, | 2008-01-02, | 2013-07-15, |
| | 2012-04-28, | 2012-06-11, | 2012-09-12, | 2017-12-11, | 2018-02-08, |
| | 2013-11-11, | 2016-09-13, | 2016-09-20, | 2018-09-30, | 2018-10-29, |
| | 2017-12-13, | 2018-01-21, | 2018-03-16, | 2021-05-20 | |
| | 2018-09-03, | 2018-10-12, | 2018-10-25, | | |
| | 2018-10-30, | 2020-01-02, | 2021-05-05, | | |
| | 2021-05-22 | | | | |

In the bias proportion series, the L2 model detects no breakpoints, suggesting that the mean level of gendered language in headlines remained relatively stable across the entire period. However, the RBF model highlights several regime shifts from as early as 2006 through to 2021. Notably, there are several breakpoints clustered around major social events, such as the #MeToo movement in late 2017 and early 2018, with changepoints detected on 2017-12-13, 2018-01-21, and 2018-03-16. These likely reflect an increased and variable focus on gendered language during this period of heightened social awareness.

In the sentiment polarity series, both models identify a significant changepoint on 2021-05-20. However, the cause of this shift is not immediately apparent, as it does not align with any major social or political events known to disproportionately affect the framing of women in headlines. Its detection across both L2 and RBF models suggests a

genuine and substantial change in the average tone of media coverage relating to women. Additionally, the RBF model identifies further sentiment breakpoints around 2008, 2012, and 2018, corresponding to periods of broader political or economic shifts that may have influenced media tone.

Overall, the RBF model provides a richer segmentation of the time series, allowing us to capture not only abrupt changes in average bias or sentiment but also subtler shifts in variance. This granularity is especially important for detecting media framing dynamics that may not manifest as large shifts in mean but still represent meaningful changes in how women are portrayed in the news.

### *5.4 Discussion*

The results of this analysis offer partial support for the alternative hypothesis $H_1$: that bias and sentiment polarity in headlines featuring women have undergone significant changes around key societal events. However, the findings also reveal important complexities, particularly in the temporal structure of the data and the modelling strategy.

A notable observation is that ARMA(1,0) forecasting models produced very narrow confidence intervals, with forecasted variances consistently near zero. This outcome is not unexpected given the nature of autoregressive models of order 1, which tend to revert quickly to the mean and assume limited persistence or memory in the series. The implication is that unless there is strong autocorrelation in the data, such models will offer predictions that are flat and conservative.

The use of ARMA(1,0) was based on extensive model selection procedures, including the minimisation of AIC and BIC scores across alternative ARMA specifications. In addition, ARCH tests were conducted on the residuals of these models to rule out volatility clustering, with results suggesting that more complex structures (e.g., GARCH or ARIMA models) were not warranted. Despite this, the question remains: could the underlying mechanisms for generating bias and polarity be inherently random? In this context, the data-generating process refers to the daily values of bias proportion and sentiment polarity in media headlines about women. If this process had meaningful temporal structure (e.g., if today's bias influenced tomorrow's), autoregressive models should be able to capture that. However, the low residual autocorrelation and the failure of our models to meaningfully predict variation post-event suggest that these series may behave similarly

to white noise, a process in which each value is effectively independent of the last. This would imply that the level of bias or polarity in one headline has little to no predictive influence on the next.

Interestingly, despite this apparent randomness, both series display signs of mean-reverting behaviour, with bias proportion fluctuating around a small positive constant and sentiment polarity remaining consistently negative. These trends appear stable across the entire dataset, with only short-term deviations observed around the #MeToo and COVID-19 event boundaries. This raises important questions about the responsiveness of editorial tone to cultural change. Even amid major societal movements, the language used to portray women in headlines appears remarkably persistent.

It is also important to consider the role of data aggregation in shaping the observed dynamics. Because sentiment polarity scores were computed as the daily mean across all relevant headlines, it is possible that strongly negative headlines were offset by more neutral or even positive ones, resulting in an average that masks underlying variation. This aggregation effect could contribute to the low autocorrelation and weak predictability of the polarity series, particularly if sentiment signals are more episodic or headline-specific. In contrast, the bias measure, being defined as a proportion and always non-negative, is less affected by such cancellation and may therefore provide a more stable signal. Future work could explore alternative aggregation strategies to assess whether richer temporal patterns can be uncovered in sentiment trends.

These findings have important societal implications. First, it suggests that although events like #MeToo may have succeeded in increasing the visibility of women's stories, they have not fully transformed how those stories are linguistically framed. If media sentiment remains predominantly negative or biased, even subtly, this may reinforce harmful gender stereotypes, regardless of increased representation. Second, the resilience of these language patterns may reflect deeper institutional inertia within journalism and media industries, where editorial practices and cultural norms lag behind social progress.

In light of these findings, future work could explore whether the lack of structural change is mirrored in other dimensions of media (e.g., imagery, narrative framing, or source attribution), or whether different genres of media (e.g., opinion pieces, tabloid versus broadsheet, or social media discourse) reveal greater responsiveness to societal movements. Long-term improvements in gender representation may require not just more

frequent mention of women in the news, but more fundamentally, a shift in the emotional, linguistic, and narrative tone through which their stories are told.

### *5.5 Validity of Results*

While this study applies robust NLP and time series techniques to explore gendered media language, certain limitations affect the validity of the findings. Internal validity may be influenced by the coverage of the custom dictionary and the sentiment model's sensitivity to contextual nuance, particularly given the condensed and sometimes ambiguous nature of headlines. Additionally, the use of pre-trained models such as RoBERTa introduces potential biases stemming from their original training corpora. External validity is constrained by the exclusive focus on U.S. news media and English-language headlines, which may limit the generalisability of findings to other cultural or linguistic contexts. Nonetheless, the study maintains strong construct validity by aligning its analytical tools with well-established theoretical concepts, such as stereotypical framing and sentiment polarity.

## 6 FURTHER RESEARCH & LIMITATIONS

Despite careful design and analysis, this study has several limitations that suggest meaningful avenues for future research.

### *6.1 Scope and Representativeness*

This analysis was conducted exclusively on English-language headlines originating from U.S.-based news sources. While this choice ensured consistency in linguistic processing and reduced cultural variation, it limits the generalisability of the findings to English-speaking and Western media environments. Future research could broaden the linguistic and geographical scope by incorporating multilingual datasets from diverse regions. This would enable cross-cultural comparisons of gender representation in media and reveal how language use around women differs across sociocultural contexts.

Moreover, this study focused exclusively on traditional news media, which, while offering structured and editorially curated content, represent only one dimension of public discourse. Social media platforms, where narratives about gender are often informal, rapidly evolving, and user-driven, may offer complementary perspectives. Although time and resource constraints restricted this analysis to news headlines, future work could integrate data from platforms such as Twitter or Reddit. Prior studies have used Twitter to examine the dynamics of the #MeToo movement, sentiment around gendered violence, and public reactions to feminist discourse (Manikonda et al., 2018). Others have analysed Reddit threads to explore the framing of gender issues in online communities and how misogynistic or supportive discourse evolves (Farrell et al., 2019). Incorporating such data would facilitate comparative sentiment and framing analyses between institutional media and everyday discourse, deepening our understanding of gender portrayal in contemporary communication ecosystems.

### *6.2 Temporal and Structural Limitations*

The dataset revealed uneven coverage over time, particularly between 2010 and 2014, when fewer headlines mentioning women were recorded. This variation complicates efforts to distinguish between genuine societal change and artifacts of data collection, such as improved archiving practices or shifts in publication volume. Future studies should aim

to normalise for data volume or implement resampling strategies to mitigate the impact of temporal imbalances. Integrating additional archival data sources could further enhance longitudinal robustness.

Moreover, the ARMA(1,0) models employed in this study, while appropriate under model selection criteria, are limited in their capacity to model abrupt or non-linear shocks. These models inherently predict values close to the mean, making them poorly suited to capturing abrupt sentiment shifts or large-scale changes following major societal events. Exploring more flexible modelling frameworks such as regime-switching models, non-linear autoregressive models, or structural time series approaches could better capture discontinuities and irregular dynamics.

### *6.3 Sentiment Aggregation Effects*

Sentiment polarity scores were computed as the daily mean across all headlines mentioning women. While this reduces noise and stabilises the series, it introduces a smoothing effect that may mask short-term variation. Strongly negative headlines could be offset by neutral or even positive ones, yielding daily averages that understate the true range of sentiment. This aggregation may partly explain the weak autocorrelation and low predictive power of the polarity series. Future research could explore alternative aggregation strategies such as median polarity, dispersion measures, or headline-level modelling to preserve more nuanced temporal dynamics.

### *6.4 Lexical Bias and Dictionary Design*

Although sentiment analysis was conducted using pretrained model-based approaches, these tools may still misclassify language that is stereotypically associated with women. For example, assertiveness or ambition may be interpreted negatively, reflecting societal biases embedded in the models' training data. While a custom gender-stereotype dictionary was used to assess bias independently, this approach introduces subjective decisions about word inclusion and stereotypical associations. Moreover, it may omit culturally specific or evolving gendered expressions. Additionally, the reliance on word-level indicators of bias fails to capture contextual subtleties. Future research could consider data-driven lexicon generation methods, such as Empath (Fast et al., 2016), which uses seed terms to expand semantic categories based on crowd-labeled text, or use dynamic word embeddings

to track how gendered meanings shift over time.

### *6.5 Debiasing and Fairness Frameworks*

This study evaluated bias and polarity in headlines but did not attempt to mitigate or correct potential bias in the underlying models or data. Future work could incorporate fairness diagnostics and debiasing techniques to improve the interpretability and ethical integrity of NLP pipelines. For example, name-swapping counterfactuals (Hall Maudslay et al., 2019) or embedding projection methods (Bolukbasi et al., 2016) could help identify and adjust for systematic gender bias. In addition, frameworks from algorithmic fairness literature (Binns, 2018) could guide more reflexive approaches to evaluating media content and the tools used to analyse it.

# 7 CONCLUSION

This thesis examined how women are portrayed in U.S. news headlines across two major societal events, the #MeToo movement and the COVID-19 pandemic, by analysing changes in gender bias and sentiment polarity. Using a custom dictionary-based metric for stereotypical language and a transformer-based sentiment classifier, daily time series were constructed and modelled using changepoint detection and ARMA forecasting.

The findings reveal that the #MeToo movement produced only modest and short-lived changes in sentiment and bias, with increased emotional volatility but no sustained shifts in linguistic framing. In contrast, the COVID-19 pandemic was associated with more pronounced and persistent fluctuations in sentiment, as well as a statistically significant increase in variance in the language used to describe women. These differences suggest that while media attention toward women can increase during major disruptions, deeper editorial conventions and cultural scripts remain largely intact.

The study contributes to research in media framing, feminist linguistics, and computational text analysis by integrating time series methods with NLP tools to uncover structural changes in language over time. Although the scope was limited to traditional news media, future research could explore how gendered narratives evolve across platforms such as Twitter or Reddit, enabling comparative analysis between institutional and user-driven discourse. Such extensions may deepen our understanding of persistent gender framing and help illuminate the broader communication dynamics that shape public perceptions of women.

# References


Balahur, A. (2013). Sentiment analysis in social media texts. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 122–128.

Barbieri, F., Espinosa Anke, L., and Camacho-Collados, J. (2022). Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, pages 258–266.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., and Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, 5(4):323–370.

Beddingfield, N. (2004). Unwritten. CD.

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*, pages 149–159.

Boczkowski, P. J. and Mitchelstein, E. (2013). *The News Gap: When the Information Preferences of the Media and the Public Diverge*. MIT Press, Cambridge, MA.

Bollegala, D., Alsuhaibani, A., Wang, T., Kit, C., and Ishizuka, M. (2016). Joint learning of word embeddings using a corpus and a semantic lexicon. In *Proceedings of AAAI*.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29.

Budanitsky, A. and Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, New York.

Cowie, L. T., Mykkeltveit, I., Larsen, G., and Kjøs, H. L. (2021). Covid-19 and women's rights: A global outlook. Technical report, University of Bergen. Accessed from the University of Bergen Institutional Repository.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 4171–4186.

Dor, D. (2003). On newspaper headlines as relevance optimizers. *Journal of Pragmatics*, 35(5):695–721.

Dragaš, M. (2012). Gender relations in daily newspaper headlines: the representation of gender inequality with respect to the media representation of women(critical discourse analysis). *Discourse (Interdiscursivity)*, 6:104.

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

Fang, X., Che, S., Mao, M., Zhang, H., Zhao, M., and Zhao, X.-Y. (2023). Bias of ai-generated content: An examination of news produced by large language models. *arXiv*, abs/2309.09825.

Farrell, T., Fernandez, M., Novotny, J., and Alani, H. (2019). Exploring misogyny across the manosphere in reddit. New York, NY, USA. Association for Computing Machinery.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., and Smith, N. A. (2015). Retrofitting word vectors to semantic lexicons. In *NAACL*.

Fast, E., Chen, B., and Bernstein, M. S. (2016). Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4647–4657.

Fast, E. and Horvitz, E. (2016). Identifying and characterizing political bias in news summarization. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1470–1480.

Fileborn, B. and Loney-Howes, R. (2019). Naming the unspeakable: Exploring the impact of metoo in australia. *Women's Studies International Forum*, 74:252–259.

Frenda, S., Gatt, A., and van der Plas, L. (2019). The journalist in the machine: Bias and fairness in automated news. In *Proceedings of the First Workshop on Media Bias in NLP*, pages 44–49.

Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Gerring, J. (2005). Causation: A unified framework for the social sciences. *Journal of Theoretical Politics*, 17(2):163–198.

Gill, R. and Orgad, S. (2022). Get unstuck!: Pandemic positivity imperatives and self-care for women. *Cultural Politics*, 18(1):114–134.

Glavaš, G. and Vulić, I. (2018). Explicit retrofitting of distributional word vectors. In *ACL*.

Haines, E. L., Deaux, K., and Lofaro, N. (2016). The times they are a-changing ... or are they not? a comparison of gender stereotypes, 1983-2014. *Psychology of Women Quarterly*, 40(3):353–363.

Hall Maudslay, E., Cotterell, R., and Augenstein, I. (2019). It's all in the name: Mitigating gender bias with name-based counterfactual data augmentation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5267–5275.

Hausman, C. and Rapson, D. (2018). Regression discontinuity in time: Considerations for empirical applications. *Annual Review of Resource Economics*, 10:533–552.

Hitti, Y., Jang, E., Moreno, I., and Pelletier, C. (2019). Proposed taxonomy for gender bias in text; a filtering methodology for the gender generalization subtype. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 8–17.

Hu, L. and Kearney, M. W. (2021). Gendered tweets: Computational text analysis of gender differences in political discussion on twitter. *Journal of Language and Social Psychology*, 40(4):482–503.

Hyndman, R. J. and Athanasopoulos, G. (2018). Arima models. In *Forecasting: Principles and Practice*, chapter 8. OTexts, 2nd edition.

Jha, R. and Mamidi, R. (2017). When does a compliment become sexist? analysis of sexism in the #metoo era. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 3154–3160.

Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics*.

Kaggle (2022). Women in headlines: Bias. https://www.kaggle.com/datasets/thedevastator/women-in-headlines-bias/data?select=word_dictionaries.csv.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355.

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

Ljung, G. M. and Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303.

Lu, K., Mardziel, P., Wu, F., Amancharla, P., and Datta, A. (2020). Gender bias in neural natural language processing. In *Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday*, pages 189–202. Springer.

Macharia, S. (2021). Who makes the news? 6th global media monitoring project report. Accessed: 2025-05-06.

MacKinlay, A. C. (1997). Event studies in economics and finance. *Journal of Economic Literature*, 35(1):13–39.

Manikonda, L., Beigi, G., Liu, H., and Kambhampati, S. (2018). Twitter for sparking a movement, reddit for sharing the moment: #metoo through the lens of social media. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 196–203.

McCombs, M. E. and Shaw, D. L. (1972). The agenda-setting function of mass media. *Public Opinion Quarterly*, 36(2):176–187.

Mendes, K., Ringrose, J., and Keller, J. (2018). metoo and the promise and pitfalls of challenging rape culture through digital feminist activism. *European Journal of Women's Studies*, 25(2):236–246.

Mills, S. (2008). *Language and Sexism*. Cambridge University Press.

Mohammad, S. M. (2020). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Mushtaq, R. (2011). Augmented dickey fuller test. *SSRN Electronic Journal*.

Purohit, S., Taneja, V., Saxena, P., et al. (2025). Analyzing two decades of media sentiments: Nlp and deep learning insights into news bias and trends. *Iran Journal of Computer Science*.

Reuters Institute for the Study of Journalism (2023). Digital news report 2023.

Santoniccolo, F., Trombetta, T., Paradiso, M. N., and Rollè, L. (2023). Gender and media representations: A review of the literature on gender stereotypes, objectification and sexualization. *International Journal of Environmental Research and Public Health*, 20(10):5770.

Seluman, I., Eguono, A., Arikenbi, G., and Aimiomode, A. (2024). Stereotypical portrayal of gender in mainstream media and its effects on societal norms: A theoretical perspective. *International Journal of Multidisciplinary Research and Growth Evaluation*, 5:743–749.

Sheng, E., Chang, K.-W., Natarajan, P., and Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. In *Proceedings of EMNLP*, pages 3407–3412.

Starkey, J. C., Koerber, A., Sternadori, M., and Pitchford, B. (2019). #metoo goes global: Media framing of silence breakers in four national settings. *Journal of Communication Inquiry*, 43(4):437–461. Original work published 2019.

Tissier, J., Gravier, C., and Habrard, A. (2017). Dict2vec: Learning word embeddings using lexical dictionaries. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Truong, C., Oudre, L., and Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, 167:107299.

Wenham, C., Smith, J., and Morgan, R. (2020). Covid-19: the gendered impacts of the outbreak. *The Lancet*, 395(10227):846–848.

World Health Organization (2020). Coronavirus disease (covid-19) outbreak. Accessed: 2025-05-09.

Xu, H., Zhang, Z., Wu, L., and Wang, C. (2019). The cinderella complex: Word embeddings reveal gender stereotypes in movies and books. *PloS ONE*, 14(11):e0225385.

## Appendix A: Artificial Intelligence Contribution Statement

Artificial Intelligence (AI) tools were used in this research solely for supportive purposes such as debugging code, formatting LaTeX documents, refining text through grammar and clarity checks, and limited brainstorming. AI was not involved in any part of the data collection, analysis, modeling decisions, interpretation of results, or formulation of conclusions. All core research contributions, including the design and motivation of analytical models, development of methodology, and critical evaluation of findings, were conducted independently by the author without the use of AI tools.

## Appendix B: Tables

Table B1: Original Gendered Language Dictionary from Source Dataset

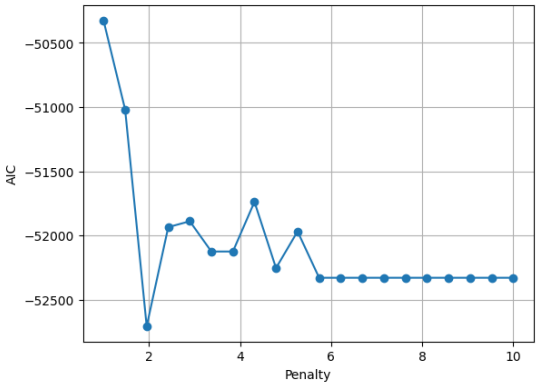
| | |
|---|---|
| **Female Gendered Words:** | klanswoman, forewoman, handywoman, airwomen, priestesses, stepmother, mum, baronesses, granddaughter, housewife, middlewomen, female, missus, hooker, womyn, bride, womanhood, assemblywomen, hoe, camerawoman, madwoman, fem, stepsister, countrywomen, sistas, queen, gal, saleswoman, clergywoman, madame, sportswomen, whore, assemblywoman, sis, duchesses, grandma, marchioness, doorwoman, ladysplain, gentleman, herself, sandwoman, goddessly, mailwomen, laywomen, countrywoman, sportswoman, councilwomen, freshwomen, fisherwomen, bogeywomen, kinswomen, spacewomen, baronetesses, momma, milkwomen, ombudswoman, camerawomen, lady, craftswoman, widow, duchess, workwomen, businesswomen, weatherwoman, alderwomen, chairwoman, moms, forewomen, freshwoman, bondswoman, ladies, mother, servicewomen, unwoman, hangwoman, congresswoman, wife, handywomen, journeywomen, watchwoman, firewomen, garbagewoman, girl, gurl, mommy, donna, diva, females, princesses, anchorwomen, spokeswomen, deaconess, superwoman, hangwomen, goddess, middlewoman, watchwomen, clergywomen, gals, harlot, postwomen, mrs, mistress, policewoman, sandwomen, mailwoman, she, wench, spacewoman, workwoman, fisherwoman, hers, daughter, women, she's, grandmas, snowwoman, chairwomen, showwomen, archduchess, heroines, margravine, stateswomen, noblewoman, cavewoman, waitress, sister, markswoman, mummy, ombudswomen, showwoman, superwomen, priestess, henchwoman, repairwoman, alderwoman, servicewoman, congresswomen, goddessliness, mama, mums, mothers, baroness, princess, aunt, wives, empresses, laydeez, craftswomen, saleswomen, maiden, weatherwomen, sistagrammer, girlfriend, widows, actresses, goddesshood, comtesse, ma'am, niece, nieces, actress, marquise, heroine, girlfriends, grandmother, womxn, archduchesses, damsel, goddesshead, repairwomen, galpal, deaconesses, stateswoman, businesswoman, queens, henchwomen, lady-romance, mom, markswomen, mamma, empress, garbagewomen, madwomen, daughters, countesses, ms, spokeswoman, noblewomen, journeywoman, kinswoman, madam, anchorwoman, councilwoman, laywoman, sisters, viscountesses, milkwoman, bondswomen, aunts, viscountess, women's, mawmaw, postwoman, womankind, woman, policewomen, countess, fiancee, godmother, mommies, firewoman, her, miss, prostitute, doorwomen, baronetess, girls, ladiez, cavewomen, bogeywoman |
| **Female Biased Words:** | affection, flatterable, moody, chatty, tenderness, maid, feel, birth, bride, caress, crazy, frumpy, nurtur, engage, breast, girly, cries, loyal, vivacious, lippy, soft, flaky, care, frail, diet, outfit, feisty, feeling, patient, shrewd, womanliness, mousey, nag, catty, emotion, curvaceous, whore, gracefully, nurtures, cook, bikini, feminazi, commit, femaleness, adorable, maternal, neurotic, toddler, ditzy, emotional, fiance, houseproud, sensitiv, divorce, ladylike, obedient, tongue, angel, quiet, modesty, attractive, beautiful, enthusiast, weak, warm, boobs, fat, sexual, kinship, man-hater, humourless, emotions, demure, trust, hysterical, inter-personal, naked, family, sexy, empathize, beyotch, flatter, depend, child, interdependent, drama, cooking, flirty, shopping, virgin, nurture, bitch, supermum, shop, whin, commitment, modest, over-sensitive, new-born, clucky, pretty, fitness, sex, round, tender, together, menstrual, barren, pregnancy, mistress, mumpreneur, nurturing, secretary, clotheshorse, supple, skinny, dress, crying, cheer, pink, conscientious, hormonal, lie, sexually, banshee, cooperate, emotiona, collab, hair, girlier, ice queen, mumsy, girlhood, diligent, newborn, witchcraft, respon, vagina, bitchfest, married, compassion, cry, graceful, dependable, single, kid, saree, empathetic, fishwife, irrational, marriage, victim, affair, lovely, wedding, man-eater, yield, judgy, helpful, period, powerless, seductive, slutty, cold, obey, afraid, judgemental, relationship, inter-dependence, share, new born, sympath, co-operate, slut, bridezilla, womanly, frigid, womanlier, honest, bitchy, domestic, kiss, romance, feminine, submissive, fit, sharin, curvy, bachelorette, considerate, inter-dependent, unladylike, fear, spinster, motherhood, gossip, sexism, gentle, marry, whitefemaleness, agree, girliest, interdependence, marrigeable, inter-dependen, body, loose, dedicated, pleasant, lip, cheat, communal, sassy, thin, tease, polite, gossipy, makeup, affectionate, seduce, beauty, tomboy, weight, kind, catfight, sensitive, granny, impatient, baby, caring, tactful, love, cougar, empath, interpersonal, nurse, easy, shame, doll, gently, understand, womanliest, sympathy, slender, committed, grace, compassionate, support, bubbly, bossy, coy, inclusive, skin, pregnant, dramatic, prude, voluptuous, schoolgirl, prostitute, trollop, shrew, lover, calm |

[1] The original dictionary was curated according to research by Xu et al. (2019).

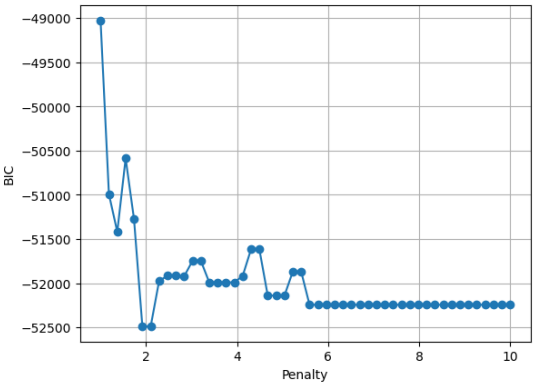
Table B2: Optimised Custom Gendered Language Dictionary

| | |
|---|---|
| **Female Gendered Words:** | aunt, hers, daughter, woman, grandmother, niece, she, mothers, her, sisters, sister, mother, girl, girlfriend, female, herself, women, wife |
| **Female Biased Words:** | womyn, donna, margravine, curvaceous, flatter, seductive, sensitiv, cavewoman, momma, widows, handywoman, skin, milkwoman, womxn, man-eater, affectionate, tactful, repairwoman, duchess, dress, hoe, mommy, firewomen, goddessly, hooker, witchcraft, mawmaw, submissive, nag, bondswomen, nurtur, servicewoman, granny, gurl, fit, feminine, forewoman, inter-dependen, cries, lip, postwoman, cavewomen, weak, sistas, man-hater, new-born, shop, weight, womanliest, craftswoman, empress, sis, archduchesses, clergywomen, mommies, banshee, clucky, diet, inter-dependence, frumpy, trust, ladysplain, councilwomen, grandmas, interdependent, depend, females, patient, hysterical, afraid, feminazi, maiden, commit, catfight, mama, viscountesses, kinswomen, saree, humourless, postwomen, bogeywomen, ombudswomen, stepmother, trollop, houseproud, nieces, gently, schoolgirl, klanswoman, superwoman, milkwomen, sistagrammer, pregnancy, interpersonal, markswoman, madame, henchwomen, fishwife, gossip, doorwoman, clotheshorse, maid, understand, kind, affection, modest, commitment, diligent, sassy, ladiez, calm, crying, toddler, superwomen, supple, outfit, frail, obey, caring, diva, tender, over-sensitive, interdependence, shopping, sandwomen, freshwomen, skinny, warm, feisty, whin, slender, gentleman, daughters, damsel, shrewd, congresswoman, inter-personal, child, empath, heroines, baroness, spacewomen, fisherwomen, harlot, moody, vagina, thin, empathize, angel, grace, caress, motherhood, saleswomen, sexually, grandma, mousey, mummy, mom, vivacious, hangwomen, baronesses, emotions, cooperate, fiancee, round, wedding, support, firewoman, ms, family, queen, nurse, businesswomen, policewoman, workwoman, garbagewomen, madwoman, pleasant, girlier, bachelorette, cold, mistress, weatherwoman, sportswoman, fiance, slutty, fear, viscountess, galpal, saleswoman, countess, nurture, attractive, bitchy, romance, helpful, honest, polite, girls, married, spacewoman, craftswomen, soft, enthusiast, fem, prostitute, domestic, priestesses, cook, irrational, care, easy, spinster, bitchfest, kinship, ladies, menstrual, love, boobs, stepsister, judgemental, sexual, fisherwoman, garbagewoman, lie, maternal, bogeywoman, obedient, compassionate, journeywoman, cheer, loyal, neurotic, baby, businesswoman, naked, divorce, beauty, frigid, mums, doll, mamma, nurtures, girlfriends, kinswoman, sympath, drama, godmother, waitress, girliest, barren, showwoman, lover, yield, new born, marry, mumpreneur, countrywoman, sandwoman, ombudswoman, princess, femaleness, respon, hormonal, hair, victim, deaconesses, queens, laywomen, womanhood, lady, mumsy, catty, journeywomen, beautiful, empathetic, baronetess, sex, feeling, bride, shrew, kid, ditzy, marrigeable, markswomen, judgy, whitefemaleness, madam, chairwoman, noblewoman, workwomen, princesses, chatty, baronetesses, single, alderwomen, showwomen, womanlier, collab, fitness, repairwomen, congresswomen, bossy, forewomen, engage, tenderness, archduchess, virgin, makeup, noblewomen, affair, sharin, watchwoman, watchwomen, housewife, weatherwomen, camerawoman, girly, bitch, quiet, spokeswomen, anchorwomen, crazy, sportswomen, aunts, countesses, anchorwoman, newborn, womankind, dependable, inter-dependent, body, committed, powerless, graceful, bubbly, marquise, adorable, considerate, laydeez, goddess, emotional, spokeswoman, pink, doorwomen, shame, unladylike, missus, gentle, snowwoman, stateswoman, clergywoman, countrywomen, co-operate, marriage, councilwoman, bridezilla, gracefully, flatterable, whore, impatient, lippy, tomboy, mailwoman, period, middlewomen, actresses, handywomen, tease, prude, freshwoman, compassion, widow, airwomen, marchioness, modesty, loose, assemblywoman, priestess, servicewomen, girlhood, emotiona, sympathy, kiss, granddaughter, emotion, relationship, deaconess, wench, lovely, pretty, supermum, cry, camerawomen, goddesshead, beyotch, flaky, womanly, madwomen, actress, bondswoman, breast, womanliness, cooking, demure, ladylike, fat, dramatic, ice queen, gals, tongue, goddessliness, empresses, dedicated, cheat, sexy, policewomen, laywoman, alderwoman, flirty, middlewoman, gossipy, mum, gal, curvy, sensitive, conscientious, mrs, heroine, moms, hangwoman, inclusive, goddesshood, secretary, she's, nurturing, communal, voluptuous, women's, wives, assemblywomen, chairwomen, pregnant, unwoman, mailwomen, duchesses, sexism, slut, feel, comtesse, seduce, share, henchwoman, miss, lady-romance, birth, bikini, ma'am, coy, together, stateswomen, cougar, agree |

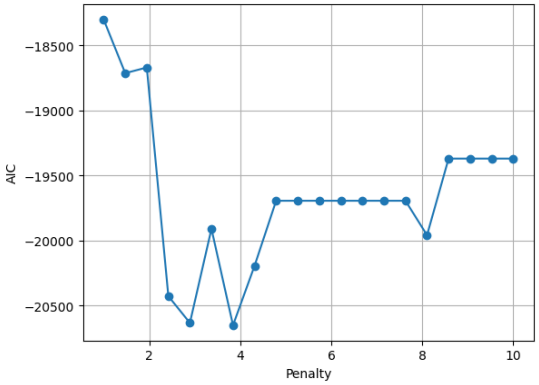
## Appendix C: Figures



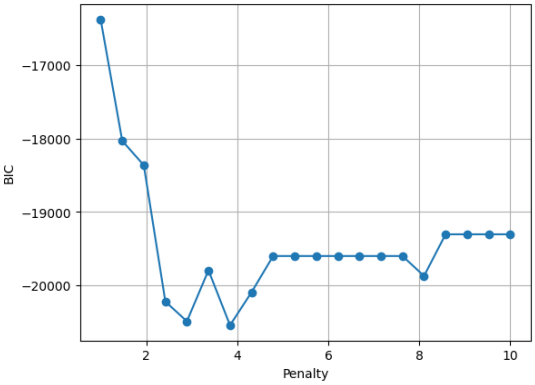

(a) AIC for bias proportion




(b) BIC for bias proportion

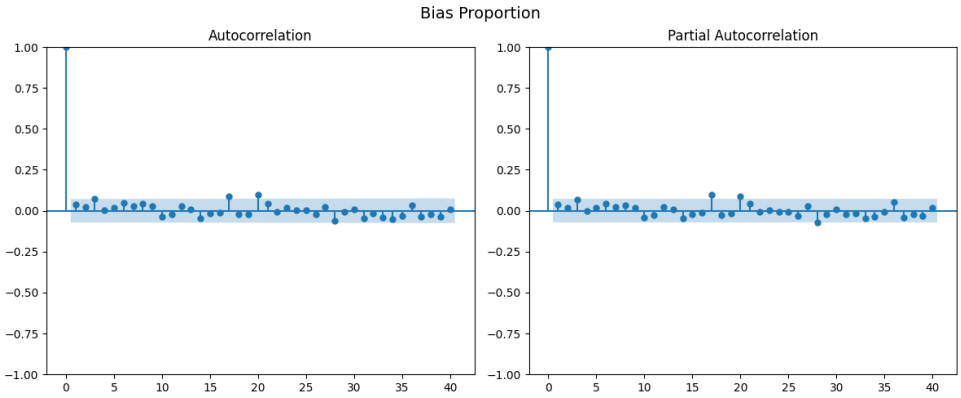



(c) AIC for sentiment polarity

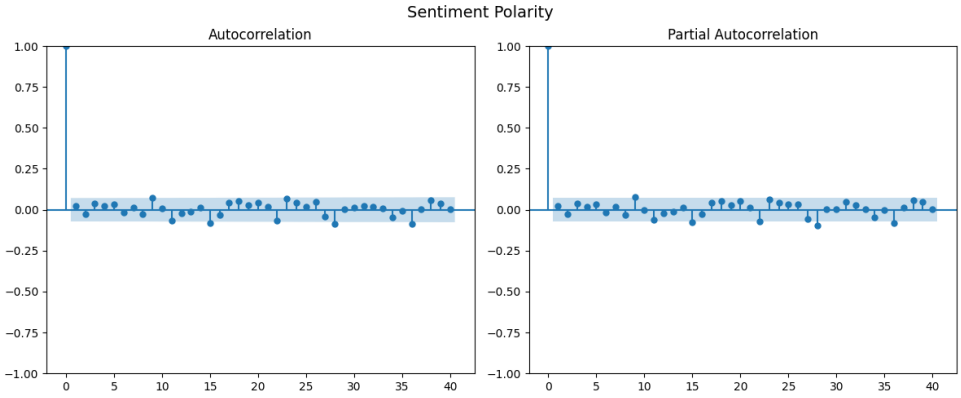



(d) BIC for sentiment polarity

Figure C1: AIC and BIC for selection of penalty term in ruptures PELT algorithm

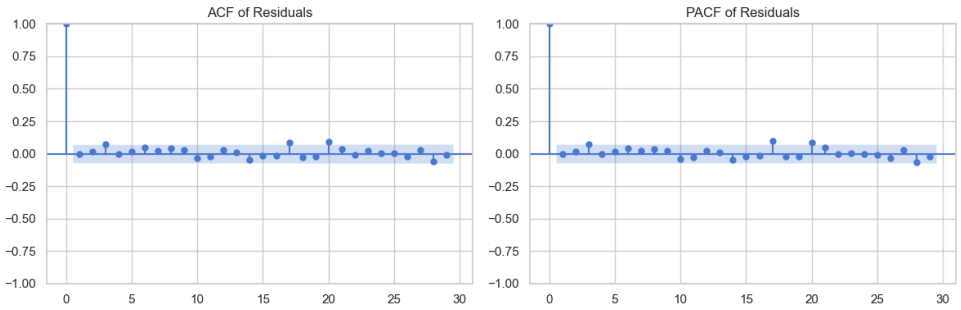



(a) Bias proportion

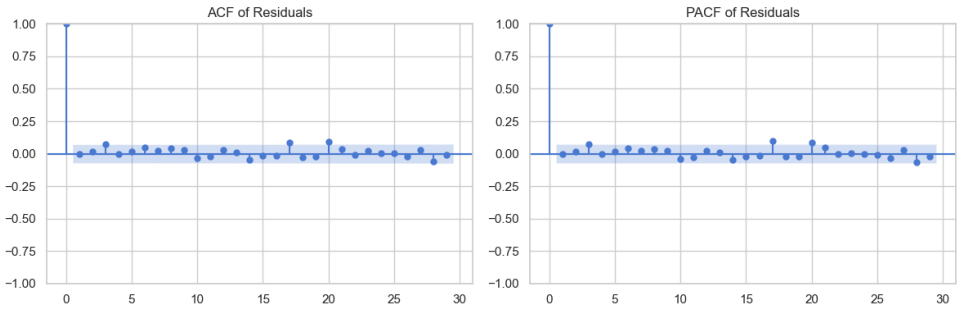



(b) Sentiment polarity

Figure C2: ACF and PACF plots for raw bias and polarity series in the pre-COVID-19 period




(a) Bias proportion




(b) Sentiment polarity

Figure C3: ACF and PACF plots for residuals of AR(1) models fitted on (a) bias and (b) polarity series in the pre-COVID-19 period