STOCKHOLM SCHOOL OF ECONOMICS
Department of Economics
5350 Master's thesis in economics
Academic year 2019–2020

A statistical analysis of gender representation in Swedish political reporting

Vera Lindén (23611)

Abstract: This thesis examines gender representation in Swedish political media using text data from Dagens Nyheter (DN), Sweden's largest daily newspaper. Descriptive statistics and regression analysis provide evidence that there are discrepancies in the amount of coverage that men and women receive; both in terms of number of articles, and article length. The existence of gendered language is studied using topic analysis and lasso-logistic regression. The results from the topic analysis show that professional titles appear more frequently in articles about men, and words related to family appear more frequently in articles about women. However, the evidence presented here does not suggest that articles about women in party leadership positions are shorter than articles about their male counterparts.

Keywords: gender, politics, quantitative text analysis

JEL: D72, J16

Supervisor: Pamela Campa
Date submitted: 18 May 2020
Date examined: 28 May 2020
Discussant: Ossian van Arkel
Examiner: Magnus Johannesson

Acknowledgements

I wish to express my gratitude towards my supervisor Pamela Campa, Stockholm School of Economics, for her invaluable guidance throughout this project. I would also like to thank Jingcheng Zhao, Per Lindström, and Anurag Dey for insightful comments and advice.

Contents

List of Tables

List of Figures

1 Introduction

In October 2019, The Economic Times, one of India's largest newspapers, published an article titled "Indian-American MIT Prof Abhijit Banerjee and wife wins Nobel in Economics". Esther Duflo is also a professor at MIT, but in this heading she was not even mentioned by name; she was only referred to as Banerjee's "wife". The title caused substantial backlash both in India and around the world, and was subsequently changed. This title is an illustrative example of how women are portrayed differently in the media, even after becoming prominent in their fields.

Women being overlooked is not unique to Indian news. According to the Global Media Monitoring Project (GMMP) (2015) women account for 24% of people heard, read about, or seen in world news. Women are less likely than men to appear in the capacity of a spokesperson, expert, or commentator (p.24). Women were most under-represented in news related to politics and government, compared to other topics discussed in the news, such as the economy; science and health; social and legal news; and crime and violence (Edström and Jacobsson, 2015, p. 27).

Adding to the existing literature on the portrayal of women in politics (Kahn and Goldenberg, 1991; Kahn, 1994; Kittilson and Fridkin, 2008; Bromander, 2012; Lühiste and Banducci, 2016), this paper studies representations of gender in Swedish political reporting; both in terms of the amount of coverage, and the language used in the coverage. Sweden is of particular interest as it is one of the most gender progressive countries in the world. According to the World Economic Forum (2019) Sweden ranks fourth in the world on gender equality. At the time of writing, Sweden has equal representation of men and women in the national Parliament (*Riksdagen*) and among cabinet ministers. However, Sweden has never had a female democratically elected head of government.

In a democracy, the media plays a central role in providing information and shaping voter perception. Evidence suggests that the visibility of women in leadership has an effect on voter perception (Beaman et al., 2009) and the career aspirations of young women (Beaman et al., 2012). This thesis contributes to the current state of knowledge by verifying the existence of a gender media gap in Sweden, and offering additional evidence on media representations of gender in an otherwise relatively gender equal context. Methodologically, this thesis is related to a growing body of research within economics which uses text as data (Wu, 2017, 2018; Gentzkow et al., 2019).

To undertake this study I have used open-source software to collect data on 8,038 articles about domestic politics from the website of Sweden's largest daily newspaper, Dagens Nyheter (DN). The articles were classified based upon the occurrence of pronouns and first names in the body of the article's text. A majority of the articles under study featured both men and women. However, articles featuring only men were three times more common than articles featuring only women: 2,508 compared to 738 articles. Articles about women who were not in a party leadership position were on average 8.7% shorter than articles about men who were not in a party leadership position. I have not found evidence supporting that articles about women in political leadership positions were shorter than articles about their male counterparts.

Using topic analysis, I have found that professional titles appeared more often in articles about men than they did in articles about women. The opposite was true for words related to family. Using a lasso-logistic regression, I determined which words were the strongest predictors for whether an article was about a male or a female. The three strongest predictors for a male article were *statsminister* ("prime minister"), *skillnader* ("differences"), and *Sverigedemokraternas* ("the Sweden Democrats""). The three strongest predictors for a female article were *make* ("husband"), *klimatminister* ("minister for the environment"), and *centerledaren* ("the leader of the centre party").

2 Literature review

This section starts by reviewing the literature on gender and language. Drawing upon scholarships in gender studies and linguistics it provides a theoretical framework for the social construct of gender and the systematic differences in language used by - and about - women compared to men. In essence, gendered language reflects an underlying power imbalance between men and women. Empirical evidence supports that the language used to describe women is often marginalising, trivialising, and objectifying.

Secondly, I consider the portrayal of gender in political media and the evidence of systematic gender differences in the news coverage of politicians. Some studies have found that female candidates receive less coverage than their male counterparts (Kahn and Goldenberg, 1991; Kahn, 1994), whereas others have not (Kittilson and Fridkin, 2008). Kittilson and Fridkin (2008) found evidence of gender stereotyped coverage with regards to the amount of attention given to different policy areas and the emphasis on different personality traits.

Thirdly, I discuss why differences in the portrayals of gender in the political media is consequential by considering the implications of differential portrayal on the efficiency of the labour market. I discuss two possible mechanisms through which differential portrayal of men and women may have an affect on career aspirations: the role model effect, and the way in which language may signal belonging.

2.1 Gender and language

For decades, scholars of gender studies have distinguished between sex (the the biological differences between men and women), and gender (a social construct). Language plays an important role in the social construction of gender identity as it "influences our perceptions and thus shapes our reality" (McGrath, 2014, p. 97). Our differing expectations of men and women are reflected in our language: "the marginality and powerlessness of women is reflected in both the ways women are expected to speak, and the ways in which women are spoken of" (Lakoff, 1973, p. 45). The language used by women is characterised by politeness, and is less assertive than male language. The language used to describe women often makes reference to their relations to men, such as marital status. As such, "men are defined in terms of what they do in the world, women in terms of the men with whom they are associated" (Lakoff, 1973, p. 64). Gender is produced and reproduced in everyday human interactions and can be understood both as "an outcome of and a rationale for various social arrangements[,] and as a means of legitimating one of the most fundamental divisions of society" (West and Zimmerman, 1987, p. 126). This division manifests itself in the form of gender gaps in income, educational attainment, health outcomes, and political power (World Economic Forum, 2019).

Empirical evidence, from a variety of literates, supports the claim that the language used - by and about - men and women differs. For example, in televised sports: "Women are often marginalized, made invisible, trivialized, infantil[ised], and reduced to sex objects" (Koivula, 1999, p. 591). Messner et al. (1993) compared verbal sports commentary and found a "hierarchy of naming" in which women were sometimes referred to as "girls" (infantilisation) but men were never referred to as "boys". Women were also more likely than men to be referred to by only their first name. The language around defeat differed between genders as well; when women lost it was typically portrayed as a result of their own shortcomings, whereas male defeat was more often framed in terms of the strength of their opponent. Messner et al. (1993) also studied the effect of

commentator gender and observed no systematic differences in the amount of gendered language used by male and female commentators respectively.

More recent empirical evidence has documented the existence of gendered language on social media and other online forums. Beltran et al. (2020) used Twitter data to show that the language used both by politicians themselves and the language used to address them was consistent with gender stereotypes. They found "evidence of gender-specific insults, and note[d] that mentions of physical appearance and infantilizing words [were] disproportionately found in text addressed to female politicians" (p.1). Similarly, Wu (2017, 2018) found evidence of gendered language on an anonymous online forum. The forum was originally intended for purely professional discussions among current and aspiring economists. Wu (2017, 2018) found that words related to a person's physical appearance and private life appeared to a disproportionate degree in discussions about women compared to men. In contrast, men were more likely to be mentioned in the context of professional and academic topics.

Gendered language may have an indirect effect on labour market outcomes. Madera et al. (2009) found that academic recommendation letters written for female applicants were more likely to mention communal adjectives, such as "kind" and "helpful", and words of socio-communal orientation, such as "wife" and "children". They also found that the prevalence of communal terms were negatively associated with experimental hireability ratings.

2.2 Portrayals of gender in political media

"By covering male and female candidates differently, the news media may influence the success of female candidates for public office" (Kahn and Goldenberg, 1991, p. 180). Some of the early work analysing portrayals of gender in political media observed that in elections for the United States' Senate female candidates received less coverage than their male counterparts (Kahn and Goldenberg, 1991; Kahn, 1994). In addition, the coverage was more likely to emphasise a female candidate's viability (her chances of getting elected) rather than her policy stance, and the viability of female candidates was generally portrayed as more negative than that of male candidates.

In contrast, a cross country study of English speaking democracies (Australia, Canada, and the United States) found that "men and women candidates [were] treated equitably in terms of amount and prominence of press attention, and the news media [did] not focus more attention on the viability or family of female candidates running for office" (Kittilson and Fridkin, 2008, p. 381). However, the news coverage reinforced gender stereotypes regarding the amount of attention given to different policy areas and the emphasis on different personality traits. The policy areas typically associated with men included economics, defence, and foreign affairs. Policy areas typically associated with women included education, health care, and equality. Personality traits disproportionally mentioned in relation to men included "strong", "competitive", "aggressive", and "independent". The list of female traits included "attractive", "passive", "emotional", and "dependent". Kittilson and Fridkin (2008) hypothesised that a greater share of women in parliament would be associated with less gender biased reporting, but did not find evidence supporting this.

On a global scale, the Global Media Monitoring Project (GMMP) has found that women are especially invisible in political media. This study is conducted during one day every fifth year. In 2015, the GMMP surveyed printed newspapers, radio, television, online newspapers, and news on Twitter in 114 countries. Compared to other topics discussed in the news, such as the economy; science and health; social and legal news; crime and violence; and celebrity, arts and sports, women were most under-represented in news related to politics and government. Only 16% of news regard-

ing politics and government featured women (Edström and Jacobsson, 2015, p. 27). Haraldsson and Wängnerud (2019) argued that this under-representation of women leads to a "false portrayal of society through a gendered lens" (p.524).

One possible explanation for the absence of women in political news may be low female representation in politics generally. In 2019, 24.9% of parliamentary seats worldwide were held by women. Similarly, 20.5% of all speakers of the house, and 6.6% of heads of state were women (UN Women, 2020).

Lühiste and Banducci (2016) offered two theoretical explanations for gender differences in political coverage: a media logic and a party logic. The media logic refers to a potential bias on behalf of news reporters, whereas the party logic refers to a potential bias in the allocation of positions within political parties. This party logic has been documented in other studies which found that women received less encouragement from a political source (Fox and Lawless, 2004), or were placed on ballots in a way which reduced their chances of being elected (Esteve-Volart and Bagues, 2012).

Lühiste and Banducci (2016) empirically studied the 2009 European Parliament election at the candidate level and found that the gender gap in coverage did not disappear after controlling for specific electoral rules and the ranking of candidates on party ballots, suggesting that mechanisms relating both to the media logic and a party logic were at play. However, they found that women were more likely to be represented in more gender equal countries.

Studies of the Swedish media have suggested that female politicians are subject to more intense scrutiny (Bromander, 2012), are cited less frequently compared to male politicians (Kroon, 2006), and that men's private life remain private to a larger extent than women's do (Wendt, 2011). It has also been argued that physical appearance is more important for women's careers than men's (Jarlbro, 2006, p. 58).

While the majority of previous research has been conducted using content analysis (which involves reading the text and recording it using a code sheet), there is precedent for more large scale studies of newspaper coverage. De Cabo et al. (2014) studied a sample of 34,235 Spanish online newspaper articles (not limited to political news) and found that women were more likely to appear in shorter news articles. They also found that female reporters were more likely to include women in their coverage.

2.3 Why is it consequential?

Haraldsson and Wängnerud (2019) argued that media representations where women are largely absent, or portrayed in a biased way, could have an effect on the political ambitions of women. A potential channel for the effect of media coverage on female career aspirations is the role model effect. Beaman et al. (2009) studied the effects of a large-scale randomised natural experiment of affirmative action in India and found that exposure to female politicians caused voters to update their preconceptions about the effectiveness of women in leadership positions. The affirmative action policy was a gender quota which randomly reserved one third of village council positions to women. A later study found positive effects of the same policy on female career aspirations, both among parents and the adolescents themselves (Beaman et al., 2012). Importantly, this was not accompanied by a negative effect on the career aspirations of adolescent boys. In addition, the gender education gap was eliminated in the villages with affirmative action, and the adolescent girls on average spent less time carrying out household chores compared to villages without the affirmative action policy.

The significance of role models has been observed to various degrees in areas beyond politics,

from corporate boards to classrooms. Kurtulus and Tomaskovic-Devey (2012) found that female representation in top management positions in American firms led to a subsequent increase of women in middle management. In contrast, a study of the Norwegian gender quota requiring that women constitute 40% of corporate boards (for publicly listed companies) found no effect beyond the direct effect on the women who were recruited to the boards (Bertrand et al., 2019). The effect of faculty gender on student outcomes has not found conclusive evidence on the effect of role models on course taking behavior (Bettinger and Long, 2005) or performance (Dee, 2007; Carrell et al., 2010; Carlana, 2019).

Another potential channel for the effect of media coverage on female career aspirations is the way in which language influences choices. Bohnet (2016) argued that the language used in job advertisements affects applicant behaviour. Empirical evidence suggests that applicants are strongly inclined to sort according to gender; both when explicitly stated (Bem and Bem, 1973; Kuhn et al., 2018), and when the preferred gender is implied through wording (Gaucher et al., 2011). Explicitly stating that ideal applicants should be male or female is typically illegal in modern democracies. However, applicants may still try to infer the preferred gender from the language used to describe the desired candidate. Gaucher et al. (2011) used a list of masculine and feminine words, such as agentic and communal attributes, to determine the gender implicitly preferred in job advertisements and found that this gendered language had an effect on perceived job appeal. Implied gender references are of particular relevance since they may have an effect on people's sorting behaviour even if this is not the intention of the author.

Sorting, the self-selection of people into different occupations, is efficient as long as it happens along desirable dimensions such as ability and motivation. But if people self-select out of certain occupations because they perceive themselves as unfit for some other (irrelevant) reason the talent pool is unnecessarily restricted (Bohnet, 2016, p. 150-152). Under the assumption that talent is equally distributed across gender "an economy that is tapping into a limited pool (men) to find its leaders must be operating inside the efficiency frontier" (Bertrand, 2018, p. 208). The economic gains of enlarging the talent pool appear to be significant. Hsieh et al. (2019) estimated that 20% to 40% of aggregated output growth in the United States between 1960 and 2010 could be attributed to an increased representation of white women, black men, and black women within high skilled occupations.

"[T]o the extent that women have internalized the traditional female gender role, they may be less attracted to leadership roles" (Eagly and Karau, 2002, p. 590). Eagly and Karau (2002) argued that women are, to a larger extent, associated with communal attributes (such as being helpful and inter-personally sensitive) whereas men are associated with agentic attributes (such as being ambitious and dominant). Leadership roles are generally associated with the same agentic attributes that are associated with masculinity. As a result, Eagly and Karau (2002) predicted that women would be disadvantaged in the pursuit of leadership positions because of the perceived dissonance between the communal attributes associated with their gender and the agentic attributes associated with the leadership role.

3 Institutional setting

This section starts by providing an overview of the Swedish general context. Sweden is widely regarded as one of the most gender equal countries in the world. However, some evidence suggest that there is still a glass ceiling: women's earnings lag behind men's, especially at the higher end of the income distribution.

Second, this section provides a background to the Swedish political landscape, the parties in the Parliament, and the party leaders. The legislator and the executive are gender balanced in the aggregate. However, Sweden has never had a female prime minister and currently none of the three largest parties have a female party leader.

Lastly, this section reviews the Swedish media landscape. The largest daily newspaper is Dagens Nyheter (DN), which is of particular relevance as it constitutes the primary data source for the empirical inquiry in this thesis. DN has a daily readership of 1,1 million people.

3.1 The Swedish general context

According to Statistics Sweden, the Swedish population was 10.3 million at the end of 2019, out of which 49.7% were women (Statistikmyndigheten SCB, 2020b). 43% of Swedes have studied at the tertiary level (Statistikmyndigheten SCB, 2020c). The share of women who have studied at the tertiary level is 49% and the corresponding number for men is 38%. During the last two decades women have accounted for between 47-48% of the labour force (The World Bank, 2019). The gender wage gap has decreased from women earning 76,6% of men's income in 2000 to 82,6% in 2018 (Statistikmyndigheten SCB, 2020a). 74% of Swedish women, and 56% of Swedish men, spend at least one hour every day on domestic chores (European Institute for Gender Equality, 2019).

Sweden is widely regarded as one of the world's most gender equal countries. The World Economic Forum (2019) ranks Sweden fourth in the world in their Global Gender Gap Report, which measures gender inequalities on economic, education, health, and political criteria. According to The Economist's Glass ceiling index, Sweden ranks second best in the world on gender equality in the workplace (The Economist, 2020). According to the World Values survey (Inglehart et al., 2014), Swedish voters are generally positive towards women in political leadership. 85.5% of Swedes said that they disagreed, or disagreed strongly, with the statement: "On the whole, men make better political leaders than women do".

However, Albrecht et al. (2003) argued that there exists a significant glass ceiling in Sweden, as measured by the increased gender wage gap in the right tail of the income distribution. The raw wage gap is approximately 50% around the ninety-fifth percentile of the income distribution. Folke and Rickne (2020) also suggested that women face significant personal costs when being promoted. Using Swedish register data from 1991 to 2012, Folke and Rickne (2020) showed that women were significantly more likely to divorce after being promoted to top political positions. Female divorce rates were above the national average whereas male divorce rates were below the national average, suggesting that family considerations can be a source of stress for women but a source of support for men.

3.2 The Swedish political landscape

Sweden does not have legislated gender quotas, but several parties have had voluntary gender quotas since the late 1970s and early 1980s. Currently, the Swedish Social Democratic Party has a zipper system which requires that men and women alternate on party lists. The Left Party has a minimum

threshold of 50% for women on party lists, and the Green Party has a 50% gender quota on party lists (Institute for Democracy and Electoral Assistance (IDEA), 2020). The Moderate Party had a quota for the 2009 election to the European Parliament mandating that the top four positions on the party list should be split evenly between men and women.

Swedish voters primarily cast their vote for a political party. In addition, there is an option of voting among the party's listed candidates by ticking the box next to their name on the ballot. If a candidate receives at least 5% of their parties votes within a constituency, he or she takes priority over the party's own list. Otherwise the distribution of seats follows the order listed by the party on their ballot (Sveriges Riksdag, 2020).

The Swedish national Parliament (*Riksdagen*) is gender balanced in the aggregate. In 2019, 47.3% of the 349 parliamentary seats were held by women (UN Women, 2019). Women accounted for 54.5% of ministerial positions. (Ministers give up their seat as member of Parliament while serving as ministers, and are replaced by another representative from the same party.) According to Statistics Sweden (2014), the share of women in the Swedish Parliament has been above 40% since the turn of the 21st century. However, Sweden has never had a female prime minister.

During the time period under study (2012-2020) there have been eight parties in the national Parliament: the Swedish Social Democratic Party (*Socialdemokraterna*), the Moderate Party (*Moderata samlingspartiet*, commonly referred to as *Moderaterna*), the Sweden Democrats (*Sverigedemokraterna*), the Centre Party (*Centerpartiet*), the Left Party (*Vänsterpartiet*), the Christian Democrats (*Kristdemokraterna*), the Liberals (*Liberalerna*), and the Green Party (*Miljöpartiet*). Swedish national elections are held on the second Sunday of September every fourth year. During the time period under study there were two elections conducted: on 14 September 2014, and 9 September 2018. Since the 2018 general election, the largest parties are the Social Democratic Party (100 seats), the Moderate Party (70 seats), and the Sweden Democrats (62 seats).

Among the three major parties, only one has had a female party leader during the time period under study: Anna Kinberg Batra was the Moderate Party leader between 2015 and 2017. (See Table 11 in Appendix 10.2 for overview of each party's respective leaders during the studied time period.) At the time of writing, four out of nine party leaders are women. The reason that there are nine party leaders for eight parties is that the Green Party always have two party leaders, one male and one female.

3.3 The Swedish media landscape

Sweden has an active media landscape characterised by a "mix of public service broadcasters, commercial legacy news media, and emerging alternative news media" (Newman et al., 2019, p. 110). Swedes consume news from a variety of channels: printed and online news sources, social media, TV, and radio (Newman et al., 2019, p. 111). The two most popular online sources are Aftonbladet online and Expressen online. 27% of Swedes pay for for online news (Newman et al., 2019, p. 111). The two most popular TV channels are SVT News (public television) and TV4 News (commercial broadcaster). Facebook is the social media platform with the widest coverage.

Government subsidies are provided to various news outlets and distributed by The Swedish Press and Broadcasting Authority (Myndigheten för press, radio och TV, 2019a). Sweden is ranked fourth best in the world regarding press freedom (Reporters without borders, 2020). Newspapers have different political stances but are not affiliated with a party. Since the 1990s, there has been an even gender representation among Swedish journalists (Jönsson, 2005).

The two largest printed daily newspapers are Dagens Nyheter (DN) and Svenska Dagbladet

(SvD). DN is Sweden's largest daily newspaper (Myndigheten för press, radio och TV, 2019b), with a daily readership of 1,127,000 between the ages of 25 and 74. DN's readers are 54% male and 46% female. DN readers spend more than the average consumer on interior decoration, wine, food, and travel (Dagens Nyheter AB, 2020a). DN self-identifies as an independent newspaper, with a liberal political orientation (Dagens Nyheter AB, 2020b).

DN relaunched it's commentary service in 2017 (Dagens Nyheter AB, 2020d), to allow readers to comment on articles. No subscription is required to make a comment but it is necessary to create an account for the commenting application (*Ifrågasätt*) using a Swedish personal number (*personnummer*) and postal code. Readers can use either their own name or an alias when leaving comments. Comments need to follow guidelines regarding relevancy and language. DN actively monitors and removes comments that are deemed inappropriate.

According to the 2015 GMMP, the representation of women in Swedish political news is significantly better than the global average. 34% of the news regarding politics and government in Sweden featured women (Edström and Jacobsson, 2015, p. 38), compared to 16% globally. This number has increased from 28% in 2005 (Gallagher, 2005, p. 122).

4 Specification of research focus

Many previous studies of mature European and European influenced democracies have focused on the news coverage during election campaigns (Kittilson and Fridkin, 2008; Lühiste and Banducci, 2016). To get a more accurate understanding of the representation of women it is important to consider periods outside of election times; as during election times it is likely that a much higher proportion of the coverage is dedicated to party leaders, who are- and have been- primarily men. Therefore, whilst representations during election times are undoubtedly important I submit that the coverage outside election periods are equally important in shaping perceptions of who belongs in the sphere of politics and power.

This thesis seeks to quantitatively verify current knowledge regarding the existence of a gender gap in the Swedish national political media. While the GMMP offers some evidence, this study is only performed on a single day, every fifth year. This sample may not be representative of the coverage over a longer period of time. Bromander (2012) also argued that there is a need for more quantitative and systematic studies of the Swedish political media.

Specifically, this thesis descriptively examines potential differences in the representations of gender in Swedish political reporting by looking at:

1. the amount of coverage,
2. the correlation between the amount of coverage and author gender,
3. the number of comments, and
4. the language used in the articles and in the comments.

5 Data

This section first describes the data collection process. Text data was extracted from DN's website using a technique called web scraping. The final dataset was a corpus (a collection of written texts) of 8,038 articles.

Second, this section outlines the methodology used for article classification. I used a list of first names and pronouns to determine whether an article should be classified as "male", "female", or "both". 2,508 articles in my corpus were classified as male, 738 articles were classified as female, and 4,209 articles were classified as both.

Lastly, this section provides an overview of additional data collection and processing I conducted. For example, these additional steps served to determine the gender of the authoring journalist, and whether an article mentioned the prime minister or party leaders.

5.1 Web scraping

Web scraping refers to the "automated gathering of data from the Internet" (Vanden Broucke and Baesens, 2018, p. 3). The advantage of having an automated agent collect and organise information from various websites is that it is less labour intensive and less prone to human error.

I collected data from DN's online news site. I scraped the articles from the section dedicated to Swedish politics (Dagens Nyheter AB, 2020c). Articles on the main page are listed in chronological order. These articles include both news articles and editorials. Figure 2 in Appendix 10.1 provides a snapshot of the main page of the section on Swedish politics. DN is written in Swedish. I have provided both the Swedish words and the English translation throughout this thesis.

I first extracted the titles and the links (urls) to each article using the *rvest* package in R. Secondly, I used Python's *BeautifulSoup* package to extract the text in the title, the date, the body, the authors (up to three author names), and the number of comments belonging to each article. To avoid manually clicking through all the pages, I used Python's Selenium package to automate the web browser. (I extracted the title in both steps as an extra identifier, in addition to the url.)

I scraped 8,964 articles dating from 19 January 2012 to 5 April 2020. 926 observations could not be scraped and were coded as missing. 388 articles in the final corpus had missing information about the author(s). Since DN relaunched it's commentary service in 2017 there were no comments before the 22 November 2017 in my sample.

5.2 Article classification

To classify the articles I removed the punctuation from the text and split the string (the character variable) into a vector of strings (i.e. a vector of individual words). I searched that vector of words to see if it contained any gender classifiers, such as pronouns and first names. For a full list of the gender classifiers please refer to Table 12 (in appendix 10.2). The selection of first names corresponds to the first names of the members of Parliament and the ministers, as of 11 March 2020. In addition, I manually looked for first names in the non-classified articles and added them to the list of classifiers.

If the text in the article's body contained only female classifiers the article was assigned to the "female" category. Similarly, if the text contained only male classifiers it was assigned to the "male" category. If neither female nor male classifiers occurred the article was classified as "neither". If both female and male classifiers occurred the article was labelled "both". Table 1 provides an overview of the raw classified data.

I only used the text from the body of the article for the classification, because the introduction often contained the name of the article author. The code required an exact match between the classifier and the words from the article. This was desirable because otherwise a name such as *Dag* might be incorrectly detected in the word *Dagens*. The code was case sensitive which prevented errors that would have occurred if the name *Dag* was matched with the word *dag* (which is Swedish for "day").

One limitation of my approach was that a person's name would not be found if it was used in the genitive case, as in denoting possession. An example of a genitive case would be if the name *Dag* was used to refer to Dag's speech (in Swedish: *Dags tal*).

Throughout this thesis, I refer to the sample of articles that were classified as either "male" or "female" as the gendered sample. I refer to the full sample as the gendered sample plus the articles classified as "both". The articles labelled "neither" was excluded from the analysis. This excluded sample mainly consisted of articles discussing political parties without mentioning specific party leaders by their (first) name, such as polling statistics. In the excluded sample, there were also short posts about live streamed debates, press conferences, and image collections.

It is important to note that since the included first names do not refer exclusively to politicians, the articles in the "female", "male", and "both" category may feature people that are not politicians themselves. For example, articles about experts, or representatives of organisations, would be included in the sample as long as DN listed these articles in their section about Swedish politics.

Category	**Number of articles**	**Average number of words**
both	4,209	533.2
female	738	301.4
male	2,508	352.1
neither	583	166.2
Total	8,038	428.8

Table 1: Sample overview

5.3 Further data collection and processing

I used the article-specific links in the gendered sample to extract the text from the comments. For practical reasons, I restricted the data collection to the first three comments, and the first 250 characters of each comment. I reclassified the articles in the "neither" category after I scraped the text from the comments. As a result, the text analysis of the comments was performed on comments belonging to 732 articles.

I manually classified the author names as male or female. In some instances, the author provided was an organisation. In the raw data there were 2,222 articles authored only by a female (or several female authors) and 3,488 articles authored only by a male (or several male authors). 437 articles had both female and male authors. 1,442 articles had an organisational author, and 111 articles had an organisational author in addition to a male or female author. For 338 articles no author names were extracted by the scraper, probably due to irregularities in the positioning of author names on the website.

During the time period under study (2012-2020) Sweden has had two prime ministers. I constructed a new variable for the occurrence of the current or former prime minister's last name

during their respective periods of tenure. I also constructed a variable for the occurrence of any party leader's last name, during their respective tenure periods. I defined the tenure period as starting from the year that he/she was elected party leader to the year that he/she stepped down. This implies that two consecutive party leaders' tenure periods would overlap in the year of transitioning form one to the other.

6 Empirical approach

This section outlines the empirical strategy used to address my research questions. The amount of coverage was measured by the number of articles (the extensive margin) and the length of these articles (the intensive margin). I investigated the relationship between author gender and the number of articles written about men and women respectively, and the length of these articles. I also compared the number of comments in the gendered sample, controlling for whether the article mentioned the prime minister or a party leader.

For measuring systematic differences in language I first looked at the occurrence of words related to different topics, including professional titles, words of socio-communal orientation (words related to family), and words related to "male" and "female" policy issues. Secondly, I used a penalised logistic regression to determine which words were the strongest predictors for whether an article belonged to the male or female category.

6.1 Amount of coverage

I measured the amount of coverage given to women and men within the Swedish political media along two dimensions: the number of articles (the extensive margin), and article length (the intensive margin). In both cases I accounted for the fact that the prime minister, and other party leaders, likely attract more attention.

6.1.1 Number of articles

The sample overview in Table 1 provides that there were more than three times as many articles in the male category than in the female category: 2,508 compared to 738. In other words, 23% of the articles in the gendered sample were about women, whereas 77% were about men. To investigate this discrepancy further I measured the average monthly number of articles in the male and female category during each year of the time period under study (2012-2020).

One would expect there to be more articles about politics, and especially about potential prime minister candidates, during election months. To account for these outliers I looked at the average monthly number of articles including and excluding election months. In the second case, I removed the election month (September) as well as the month before, and the month after the election.

To understand whether the observed gender gap in Table 1 can be attributed to the fraction of men and women in leadership positions (or whether it is reflective of the political news in general) I split the sample depending on whether a party leader was mentioned, and whether the prime minister was mentioned.

In addition to showing the monthly average of male and female articles in each year I estimated the following article level equations:

$$female_i = \beta_0 + \beta_1 election\ month_i + \epsilon_i \tag{1}$$

$$female_i = \beta_0 + \beta_1 prime\ minister_i + \beta_2 party\ leader_i + \epsilon_i \tag{2}$$

Where $female_i$ is an indicator variable taking the value of 1 if the article was classified as "female", according to the classification in section 5.2. $election\ month_i$ is an indicator variable taking the value of 1 if the article was written during August, September, or October of an election year. $prime\ minister_i$ is an indicator variable for whether the current or former prime minister

was mentioned in article i (during their time in office), and $party\ leader_i$ is an indicator variable for whether any party leader was mentioned in article i (during their tenure).

6.1.2 Article length

To measure gender differences in article length (Wu, 2017; De Cabo et al., 2014) I estimated the following article level equation:

$$\begin{aligned} log(word\ count_i) = \beta_0 + \beta_1 female_i + \beta_2 prime\ minister_i \\ +\beta_3 party\ leader_i + \beta_4 female_i * party\ leader_i + \epsilon_i \end{aligned} \tag{3}$$

Where $word\ count_i$ is the number of words in article i. Since the length of the articles did not follow a normal distribution I used the log length. $female_i$ is an indicator variable taking the value of 1 if the article was classified as "female", according to the classification in section 5.2. $prime\ minister_i$ is an indicator variable for whether the current or former prime minister was mentioned in article i (during their time in office), and $party\ leader_i$ is an indicator variable for whether any party leader was mentioned in article i (during their tenure).

Equation 3 does not distinguish between being the leader of small party and a large party, and thus assumes that the effect of party leadership on article length is constant. This is a strong assumption, especially during election years since the leaders of large parties are more likely candidates for the prime minister post.

To account for the fact that the size of the party may have a significant impact on the amount of attention that their party leader receives I restricted the sample to those parties that have had both a male and a female party leader between 2012 and 2020: the Moderate Party, the Green Party, the Liberals, and the Christian Democrats. I estimated the following equation for articles about each party individually:

$$\begin{aligned} log(word\ count_i) = \beta_0 + \beta_1 female_i + \beta_2 election\ year_i \\ +\beta_3 female_i * election\ year_i + \epsilon_i \end{aligned} \tag{4}$$

Where $word\ count_i$ is the number of words in article i; $female_i$ is an indicator variable for the gender of the person discussed in the article; and $election\ year_i$ in an indicator variable for whether the article was published during an election year. In the case of the Moderate Party I also included an indicator variable for prime minister, since Fredrik Reinfeldt held the position of prime minister until 2014. Equation 4 provides an opportunity to study whether there are any gender differences in coverage of individuals in the same position.

6.2 Author gender

To test if the author gender correlates with any observed gender gaps in the amount of coverage I estimated the following article level equations:

$$female_i = \beta_0 + \beta_1 female\ author_i + \epsilon_i \tag{5}$$

$$\begin{aligned} log(word\ count_i) = \beta_0 + \beta_1 female_i + \beta_2 female\ author_i \\ +\beta_3 female_i * female\ author_i + \epsilon_i \end{aligned} \tag{6}$$

Where $female_i$ is an indicator variable for the classification of article i; $female\ author_i$ is an indicator variable for author gender; and $word\ count_i$ is the number of words in article i.

Equations 5 and 6 assume that there are no systematic differences between male and female authors with regards to the types of news that they cover. For example, I have assumed that female authors are not more likely to write about the prime minister, compared to male authors.

6.3 Number of comments

To understand whether male and female articles received varying amounts of comments I estimated the following article level equation:

$$\begin{aligned} number\ of\ comments_i = \beta_0 + \beta_1 female_i + \beta_2 prime\ minister_i \\ +\beta_3 party\ leader_i + \beta_4 female_i * party\ leader_i + \epsilon_i \end{aligned} \tag{7}$$

Where $number\ of\ comments_i$ is the number of comments under article i; $female_i$ is an indicator variable for the classification of article i; $prime\ minister_i$ is an indicator variable for whether the current or former prime minister was mentioned in article i; and $party\ leader_i$ is an indicator variable for whether a party leader was mentioned in article i.

There were not enough comments to filter by party, so an equivalent of Equation 4 was not feasible for the comments.

6.4 Language

For measuring systematic differences in language I followed the procedures outlined in Wu (2017, 2018). I first looked at the occurrence of words related to different topics, including professional titles, words of socio-communal orientation (words related to family), and words related to "male" and "female" policy issues. Secondly, I used a penalised logistic regression to determine which words were the strongest predictors for whether an article belonged to the male or female category.

It should be noted that any family words that occurred in the text did not necessarily refer to the person's own family. I have assumed that nothing apart from gender determined the occurrence of the topics mentioned above. For example, I have assumed that the prime minister was not more commonly discussed in the context of certain policy issues.

6.4.1 Topic analysis

To study the occurrence of particular topics in male and female articles I estimated the following two equations, similarly to the approach in Wu (2017):

$$Topic_{i,j} = \beta_0 + \beta_1 female_i + \epsilon_i \tag{8}$$

$$D_{i,j} = \beta_0 + \beta_1 female_i + \epsilon_i \tag{9}$$

$$Topic_j \in \{professional\ titles, family\ words, "male"\ policy\ issues, "female"\ policy\ issues\}$$

Where $Topic_{i,j}$ is the percentage share of words belonging to topic j in article i; $D_{i,j}$ is an indicator variable for the occurrence of topic j in article i; and $female_i$ is a gender indicator variable.

The selection of topics was done manually after retrieving a list of the most common words in the full sample. To generate the list I first tokenised the text to get a "bag-of-words" representation of the text. This involved splitting the sentences of the articles into individual words, which were then treated as independent features of the text. This implies that information related to inter-word dependence and their sequential order were no longer considered. After tokenising the text I reduced the words to their stems, and ordered the words depending on their frequency. I restricted my analysis to words that appeared at least 100 times in the full sample of articles.

I analysed the list of words and grouped them into categories related to the topics discussed in previous literature. I did not find words referring to physical appearance, agentic traits, or communal traits. I chose to estimate the prevalence of professional titles, family words, "male" policy issues, and "female" policy issues. Among the professional titles, I excluded *biträdande* ("vice") because this word needs to be paired with another word to make a title, and I wanted to avoid any potential double counting of titles (for example by counting vice president as two titles instead of one). For categorising the "male" and "female" policy issues I followed the approach suggested by (Kittilson and Fridkin, 2008, p. 383). The comprehensive list of words belonging to each topic, and their translation, can be found in Table 14 (in Appendix 10.2).

6.4.2 Penalised logistic regression

There is undeniably an element of subjectivity in the topic analysis outlined above, both in terms of the choice of topics, and the words assigned to each topic. The results depend on the assumptions made regarding what constitutes "male" and "female" policy issues. I followed the classification suggested by Kittilson and Fridkin (2008) but acknowledge that there may be justified objections to this classification. Therefore, as a complementary method, I used a penalised logistic regression as in Wu (2017, 2018).

The posterior probability that an articles is classified as female, given the language used in the article, is given by (Wu, 2017, p. 39):

$$P(Female_i = 1|W_i) = \frac{e^{\theta_0 + W_i^T \theta_1}}{1 + e^{\theta_0 + W_i^T \theta_1}} \tag{10}$$

$$P(Female_i = 0|W_i) = \frac{1}{1 + e^{\theta_0 + W_i^T \theta_1}} \tag{11}$$

Where W_i is a vector of word frequencies. The likelihood for each observation is given by:

$$P(Female_i|W_i) = P(Female_i = 1|W_i)^{Female_i} * P(Female_i = 0|W_i)^{1-Female_i} \tag{12}$$

Under the assumption of independent observations, the log likelihood of N observations is given by:

$$l_N(\theta) = ln(\Pi_{i=1}^N P(Female_i|W_i)) = \sum_{i=1}^{N} Female_i * (\theta_0 + W_i^T \theta_1) - ln(1 + e^{\theta_0 + W_i^T \theta_1}) \tag{13}$$

The objective function for estimating the coefficients θ is:

$$\hat{\theta_\lambda} = argmin_\theta(-l_N(\theta)) + \lambda||\theta||_1 \tag{14}$$

where $||\theta||_1 = \sum_{j \geq 1} |\theta_j|$, i.e. the lasso penalty term. λ is the weight assigned to the penalty term (Hastie and Qian, 2016). Penalised linear models, such as the penalised logistic regression, is one of the most widely used methods for document classification with high-dimensional data (Gentzkow et al., 2019, p. 542). In this case high-dimensionality refers to where there are many possible predictors (words, or tokens) relative to the number of observations (articles). The penalised logistic regression is used for predictions, and hence will produce a list of words that occur disproportionately often for each gender.

The reason that the penalised linear regression performs well in high-dimensional settings is that the penalty term shrinks coefficients of poor predictors toward zero. This generally produces better predictions on the test set (the data not used for training the model), but it also causes the estimated coefficients to be biased (Wu, 2017, p. 8). Because the resulting coefficients are biased I did not interpret the coefficients beyond their relative magnitude.

There are a three possible penalties, each producing a different regression: the l_1-norm produces the lasso regression, the l_2-norm produces the ridge regression, and a combination produces an elastic net. I used the lasso penalty because it shrinks the coefficients of poor predictors to zero. The ridge regression would shrink coefficients close to zero, but never all the way to zero. The elastic net is a combination of the two; it would shrink some coefficients to zero, but not as many as the lasso.

Lasso stands for 'least absolute shrinkage and selection operator' (Tibshirani, 1996). The appropriate way of calculating standard errors for lasso-penalised regressions is still debated (Kyung et al., 2010). The calculation of standard errors for the lasso-logistic regression is beyond the scope of this thesis.

To perform the lasso-logistic regression I transformed all words to lower case letters and removed gender classifiers, last names, numbers, years, and rare words. I set the cutoff for rare words as those occurring less than 10 times across all the articles in the gendered sample. I represented the data in a document-term-matrix with the articles (the unit of observation) in the rows and individual words in the columns. Because the number of words were large relative to the number of articles this became a sparse matrix.

I split the data randomly into a 75% training and a 25% test set (Wu, 2018, p. 176). The training set was used to fit the lasso-logistic regression, whereas the test set was used to evaluate the accuracy of the predictions from the trained model. To determine the appropriate value of λ I used 5-fold cross-validation (Wu, 2018, p. 176). K-fold cross-validation is a resampling method which splits the training data into k folds of approximately equal size, and sequentially fits the model and calculates a measure of goodness-of-fit using each fold as a validation set (James et al., 2013, p. 181). I chose the value for λ which minimised the binomial deviance across these k (in this case five) folds.

Accuracy was calculated as the share of correctly classified articles. I made no distinction between different types of miss-classifications: predicting that an article belonged to the female category while it actually belonged to the male category was weighted equally to the opposite error. To understand how often the selected predictors appeared in the overall corpus I also included the frequencies of these words in the gendered sample.

7 Results

This sections presents the results of the empirical analysis. Section 7.1 presents the results for gender differences in the amount of coverage in the gendered sample (3,246 articles). Section 7.2 examines the correlation between author gender and the amount of coverage, using those articles in the gendered sample which were authored by either only female or only male journalists (2,243 articles). Section 7.3 investigates differences in the number of comments written by readers in the gendered sample of articles. This analysis was restricted to the gendered sample of articles published since September 2017, when the comment functionality was re-introduced (1,286 articles). Section 7.4 presents the results from the language analysis of articles in the gendered sample, and the comments to these articles.

7.1 Amount of coverage

The amount of coverage in the gendered sample was measured along two dimensions: the average monthly number of articles (the extensive margin), and article length (the intensive margin). This analysis was performed on the gendered sample (3,246 articles).

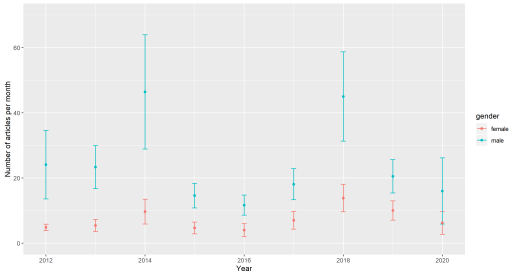
7.1.1 Number of articles

Figure 1 shows the monthly average number of articles in the male and female category between 19 January 2012 and 5 April 2020. The bars represent the 95% confidence interval. Figure 1a shows that the number of articles in the male category was consistently higher for all years except 2020, which is likely attributable to fewer observations from the present year. In Figure 1b the election months are excluded, as well as the month before and after the election (August, September, and October of 2014 and 2018). Excluding the election months reduced the gap in average monthly number of articles, but it was still persistent.

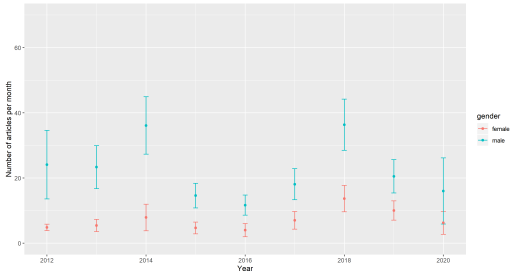
Figure 1c and 1d show the distribution of articles that either do or do not mention a party leader (including the prime minister). Figure 1d thus includes articles mentioning ministers other than the prime minister. Figure 1e represent the sub-sample of articles that mention a party leader who was not the prime minister. Figure 1f represent all articles except those articles which mentioned a prime minister.

Male articles were consistently more common, but the gender gap seems to have decreased somewhat over time. Removing the prime minister from the sample of articles about party leaders markedly reduced the gender gap, even if it did not completely disappear. Similarly, the gender gap in articles that did not mention party leaders or the prime minister exhibited a reduced gender gap.

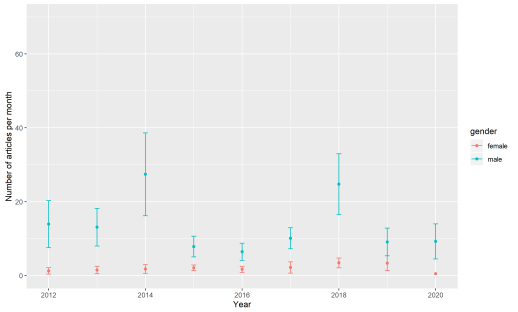
The regression results for Equation 1 and 2 are provided in Table 15 of Appendix 10.2. Column (1) provides that the share of female articles was 7.4 percentage points lower during election months, significant at the 1% level. Column (2) provides that the share of female articles was 12.2 percentage points lower if the article mentioned a party leader who was not the prime minister, significant at the 1% level.



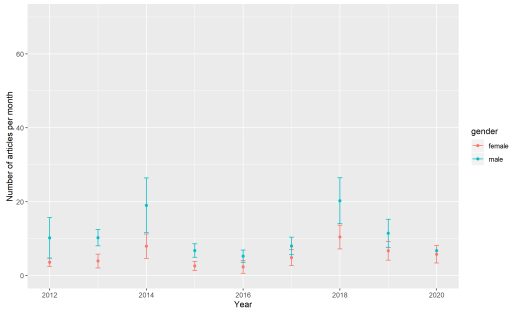
(a) All articles in gendered sample



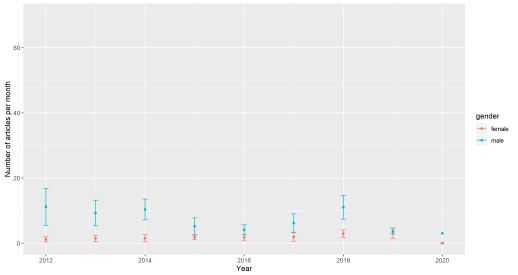
(b) Election months excluded



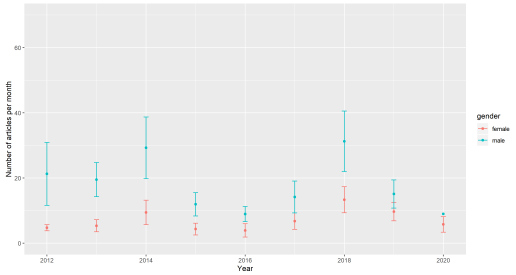
(c) Only articles mentioning party leader or prime minister



(d) Articles not mentioning party leaders or the prime minister



(e) Only articles mentioning party leader, excluding the prime minister



(f) All articles except those mentioning the prime minister

Figure 1: Average monthly number of articles with 95% confidence intervals

7.1.2 Article length

The average length of articles in the gendered sample was 340.6 words. Table 2 shows the results for the specifications in Equation 3 and 4. Column (1) of Table 2 provides that female articles which did not mention the prime minister or a party leader were on average 8.7% shorter than male articles which did not mention the prime minister or a party leader, significant at the 5% level. Articles about the prime minister were on average 22.8% longer than male articles which did not mention the prime minister or a party leader, significant at the 1% level. Articles about male party leaders were on average 10.6% longer than male articles which did not mention the prime minister or a party leader, significant at the 1% level. The coefficient for the interaction term between the female indicator variable and the party leader indicator variable is not significant, suggesting that party leaders received the same amount of attention (in terms of article length) regardless of their gender.

The regressions in columns (2)-(5) support the result that female party leaders did not receive less coverage, in terms of article length, than their male counterparts. These are regressions for the sub-sample of articles which discussed leaders of a specific party: the Moderate Party, the Green Party, the Liberals, and the Christian Democrats. The coefficient for the female indicator variable is non-significant in all regressions apart from column (5). In column (5) the coefficient is positive, which suggests that articles about the female party leader were on average longer. However, the sample size for the regression in column (5) was merely 102 observations.

Table 2: Results for article length (gendered sample)

	Dependent variable:				
	log(word count)				
	(1)	(2)	(3)	(4)	(5)
Female article	-0.083**	-0.125	-0.191*	0.093	0.325**
	(0.036)	(0.116)	(0.112)	(0.191)	(0.148)
Prime minister	0.205***	0.172***			
	(0.036)	(0.063)			
Party leader	0.101***				
	(0.032)				
Female * Party leader	-0.028				
	(0.064)				
Election year		0.174**	0.313**	0.438***	0.427
		(0.071)	(0.145)	(0.120)	(0.309)
Female article * Election year		-0.377	-0.141		-0.633
		(0.436)	(0.236)		(0.390)
Observations	3,246	428	165	203	102
R^2	0.036	0.042	0.056	0.063	0.059
Adjusted R^2	0.035	0.033	0.038	0.054	0.030
Residual Std. Error	0.687 (df = 3241)	0.591 (df = 423)	0.576 (df = 161)	0.683 (df = 200)	0.601 (df = 98)
F Statistic	30.003*** (df = 4; 3241)	4.629*** (df = 4; 423)	3.183** (df = 3; 161)	6.729*** (df = 2; 200)	2.045 (df = 3; 98)
Note:					*p<0.1; **p<0.05; ***p<0.01

7.2 Author gender

Table 3 presents the estimations of Equations 5 and 6. The sample was restricted to the gendered sample of articles which were authored by either only female or only male journalists (2,243 articles).

Column (1) of Table 3 provides that female journalists were not more or less likely to write articles about women. In column (2), the coefficient for the interaction term is insignificant, suggesting that gender differences in article length did not correlate with the gender of the journalist writing the article.

Table 3: Results for author gender

	Dependent variable:	
	Female article	log(word count)
	(1)	(2)
Female author	0.029	0.099***
	(0.018)	(0.031)
Female article		-0.103**
		(0.041)
Female article * Female author		-0.035
		(0.065)
Observations	2,243	2,243
R^2	0.001	0.011
Adjusted R^2	0.001	0.009
Residual Std. Error	0.414 (df = 2241)	0.626 (df = 2239)
F Statistic	2.609 (df = 1; 2241)	7.936*** (df = 3; 2239)
Note:		*p<0.1; **p<0.05; ***p<0.01

7.3 Number of comments

Table 4 shows the results from estimating Equation 7. Since it was not possible to comment on articles in DN before September 2017 I removed any observations before that. The resulting sample size was 1,286 articles, 779 of which had at least one comment.

Female articles, which did not mention the prime minister or a party leader, received on average 4.5 fewer comments, compared to male articles which did not mention the prime minister or a party leader, significant at the 5% level. Articles mentioning the prime minister received on average 8.0 more comments than male articles which did not mention the prime minister or a party leader, significant at the 1% level. There was no significant correlation between the number of comments and whether the article mentioned a party leader, male or female.

Table 4: Results for number of comments

	Dependent variable:
	Number of comments
Female article	-4.486**
	(1.907)
Prime minister	8.022***
	(2.093)
Party leader	-0.686
	(1.949)
Female * Party leader	3.948
	(3.459)
Observations	1,286
R^2	0.026
Adjusted R^2	0.023
Residual Std. Error	24.270 (df = 1281)
F Statistic	8.511*** (df = 4; 1281)
Note:	*p<0.1; **p<0.05; ***p<0.01

7.4 Language

This section presents the results from the topic analysis and the lasso-logistic regression on the language in the gendered sample (3,246 articles, and comments belonging to 732 of those articles). Section 7.4.1 presents the topic analysis on the article text. Section 7.4.2 presents the results from the lasso-logistic regression on the article text, and section 7.4.3 presents the results from the lasso-logistic regression on the text in the comments.

7.4.1 Topic analysis

Table 5 corresponds to Equation 8. Column (1) of Table 5 suggests that articles about women featured fewer professional titles compared to articles about men. 0.56% of the words in the male articles were professional titles. The share of words in female articles which were professional titles was 0.06 percentage points lower, significant at the 5% level.

Column (2) of Table 5 suggests that articles about women featured more family words than articles about men. 0.04% of the words in the male articles were related to family. The share of words in the female articles which were related to family was 0.04 percentage points higher, significant at the 1% level.

Column (3) of Table 5 suggests that the share of words related to "male" policy issues was higher in articles about women. The share of words related to "male" policy issues was 0.17% in the male sample, and 0.04 percentage points higher in the female sample, significant at the 1% level. The results in column (4) suggest that there was no significant difference between the share of words related to "female" policy issues. The share of words that were related to "female" policy issues was 0.17% in the male sample, with no significant difference for the female sample.

Table 5: Results for topic analysis

	Dependent variable: share (%) of words			
	professional titles	family words	"male" issues	"female" issues
	(1)	(2)	(3)	(4)
Female article	-0.063**	0.042***	0.039***	0.009
	(0.031)	(0.008)	(0.014)	(0.015)
Constant	0.559***	0.043***	0.168***	0.165***
	(0.015)	(0.004)	(0.006)	(0.007)
Observations	3,238	3,238	3,238	3,238
R^2	0.001	0.009	0.003	0.0001
Adjusted R^2	0.001	0.008	0.002	-0.0002
Residual Std. Error (df = 3236)	0.748	0.187	0.323	0.352
F Statistic (df = 1; 3236)	4.077**	28.019***	8.123***	0.388
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 6 corresponds to Equation 9. The regression in column (1) of Table 6 suggests a significant difference in the share of male and female articles that included professional titles. 77.5% of the male

articles mentioned at least one professional title, whereas 66.7% of the female articles mentioned at least one professional title, significant at the 1% level.

The regression in column (2) of Table 6 suggests a significant difference in the share of male and female articles that included words related to family. 10.9% of the male articles mentioned at least one word related to family, and 17.1% of the female articles mentioned at least one word related to family, significant at the 1% level.

Column (3)-(4) show no statistically significant differences between the shares of articles that mentioned "male" and "female" policy issues.

Table 6: Results for topic analysis

	Dependent variable: indicator variable for			
	professional titles	family words	"male" issues	"female" issues
	(1)	(2)	(3)	(4)
Female article	-0.108***	0.062***	0.019	0.0001
	(0.018)	(0.014)	(0.020)	(0.019)
Constant	0.775***	0.109***	0.362***	0.306***
	(0.009)	(0.007)	(0.010)	(0.009)
Observations	3,238	3,238	3,238	3,238
R^2	0.011	0.006	0.0003	0.000
Adjusted R^2	0.011	0.006	-0.00002	-0.0003
Residual Std. Error (df = 3236)	0.430	0.328	0.482	0.461
F Statistic (df = 1; 3236)	36.027***	20.259***	0.920	0.00002
Note:				*p<0.1; **p<0.05; ***p<0.01

7.4.2 Lasso-logistic regression: article text

From the vocabulary used in the article text I removed gender classifiers, rare words (words occurring less than 10 times in the sample), last names, and numbers. The remaining 7,231 words were used to train the model. For the optimal value of λ, the lasso-logistic regression selected 248 words as predictors for gender. Figure 3a in Appendix 10.1 shows the number of predictors selected for different values of λ. The left vertical line corresponds to the λ which minimises the binomial deviance.

Table 7 and Table 8 provide the ten strongest predictors (excluding the intercept) for female and male articles respectively. The strongest 40 predictors are provided in Table 16 and Table 17 in Appendix 10.2. The number of times each word occurred in the gendered sample is provided in the last column, denoted *n*.

The strongest predictor for a female article was the word *make* ("husband"), which appeared eleven times in the gendered sample. Six out of ten words made reference to ministerial posts or a party leadership position. The remaining three words were *gymnasie* ("high school"), *migrationspolitisk* ("migration politics"), and *religion* ("religion").

The strongest predictor for a male article (apart from the intercept) was *statsminister* ("prime minister"), which was mentioned 768 times in the gendered sample. There were three references to

specific parties among the top ten predictors. The remaining three words were *skillnader* ("differences"), *heta* ("called", or "hot"), *professor* ("professor"), and *förre* ("former").

In general, male predictors occurred more frequently than the female predictors. The intercept had the same sign as the coefficients for the male predictors since the majority of articles belonged to the male category.

Predictor (Swe)	Translation (Eng)	Coefficient	n
make	husband	1.71	11
klimatminister	minister for the environment	1.65	11
centerledaren	the leader of the centre party	1.53	41
gymnasie	high school	1.51	10
jämställdhetsminister	minister for gender equality	1.41	34
arbetsmarknadsminister	minister of labour	1.27	36
socialförsäkringsminister	minister for social security	1.25	10
miljöminister	minister for the environment	1.15	26
migrationspolitisk	migration politics	1.08	12
religion	religion	0.99	10

Table 7: Strongest predictors for female article

Predictor (Swe)	Translation (Eng)	Coefficient	n
statsminister	prime minister	-0.45	768
skillnader	differences	-0.38	62
sverigedemokraternas	the Sweden Democrats'	-0.37	516
heta	called, or hot	-0.32	18
professor	professor	-0.31	107
finanserna	finance	-0.22	35
statsministern	the prime minister	-0.22	385
förre	former	-0.21	83
vänsterledaren	the left party leader	-0.17	32
sverigedemokraterna	the Sweden Democrats	-0.16	1,275

Table 8: Strongest predictors for male article

For evaluating the model's accuracy in predicting article gender in the test set I chose a threshold of -0.55, because this maximised accuracy on the test set. The threshold is the value above which the articles were predicted to belong to the female category. The accuracy rate was 80.6% on the test set for a threshold value of -0.55.

7.4.3 Lasso-logistic regression: comment text

The number of possible predictors in the comment text was 585 words. This was considerably smaller than in the article text since the comments constituted a smaller corpus. For a value of λ which minimised the binomial deviance, 68 predictors were selected. Figure 3b in Appendix 10.1 shows the number of predictors selected for different values of λ.

Table 9 and Table 10 provide the ten strongest predictors (excluding the intercept) for comments belonging to female and male articles respectively. All predictors for each category are shown in Table 18 and Table 19 of Appendix 10.2.

The strongest predictor for comments on a female article was the word *kvinna* ("woman"), which appeared eleven times in the comments belonging to the gendered articles. The strongest predictor for comments on a male article was the word *valet* ("the election", or "the choice"), which appeared 40 times.

The model performed poorly on the test set. It had a constant accuracy rate of 76.5% for all thresholds. In addition (as shown in Figure 3b of Appendix 10.1), the accuracy in the training set did not decrease as the number of predictors approached zero. This implies that even the strongest predictors had poor predictive power.

Predictor (Swe)	Translation (Eng)	Coefficient	n
kvinna	woman	1.03	11
behov	needs	0.73	12
skatt	tax	0.68	19
medborgare	citizen	0.65	23
sådant	such	0.62	18
senare	later	0.58	10
innebär	means	0.50	19
landsting	county	0.48	11
barn	child, or children	0.46	21
skatter	taxes	0.45	13

Table 9: Strongest predictors for female

Predictor (Swe)	Translation (Eng)	Coefficient	n
valet	the election	-0.55	40
partierna	the parties	-0.46	32
emot	against	-0.43	25
naturligtvis	naturally	-0.40	16
förstå	understand	-0.38	21
liten	small	-0.35	14
haft	had	-0.30	22
resurser	resources	-0.27	15
partiet	the party	-0.21	42
ingen	no one	-0.16	91

Table 10: Strongest predictors for male

8 Discussion

This section provides an interpretation of the findings presented above and their possible implications. In comparing the article length and the number of comments, I would consider article length to be more important, primarily because readers who self-select into commenting are unlikely to be representative of the population at large. I also discuss some important limitations of my study and provide suggestions for future research.

Figure 1 and Table 2 suggest that the single most important factor in determining the amount of coverage that men and women get in political media is who holds the prime minister post. This is hardly surprising, and suggests that the persistent gender gap in media representation may be closed, or even reversed, if and when Sweden elects a female prime minister.

The results in Table 2 also suggest that women in party leadership positions do not receive less coverage (in terms of article length), compared to men in the same position. This finding is contradictory to the media logic proposed by Lühiste and Banducci (2016), according to which gender disparities in media representation is a result of a journalistic bias. Rather than a biased media, the gender gap in media representation may be a reflection of real words imbalances in political power between men and women. Despite having equal representation in Parliament this finding would suggest that Sweden has not yet closed the gender gap in politics.

In contrast to previous research (De Cabo et al., 2014) I did not find any significant correlation between author gender and the gender of the person(s) in the article. Neither did I find any significant correlation between author gender and the gender gap in article length. A possible reason for this inconsistency is a difference in research scope between the study conducted by De Cabo et al. (2014) and this thesis. De Cabo et al. (2014) conducted a broader study of the Spanish media and did not limit the scope to political news.

The results in Table 4 provide that articles about the prime minister received more comments. Articles about women who did not hold party leadership positions received less comments. This is consistent with the results for article length provided in Table 2. Since ministerial titles were among the strongest predictors for female articles, I hypothesise that a substantial proportion of the articles about women that do not hold party leadership positions constitute coverage of cabinet ministers. If this is the case, the observation that articles about women were shorter and had fewer comments is somewhat surprising, especially given that half the ministerial positions are currently held by women.

The first conclusion from the language analysis is that I did not find any words referring to physical appearance, communal traits, or agentic traits in the article text or the comments. The absence of this type of gendered language in the article text is consistent with the findings of Kittilson and Fridkin (2008). The fact that this language is also absent from the comments may seem somewhat contradictory to previous research (Wu, 2017; Beltran et al., 2020). However, since people are required to provide their personal number to be able to comment, it is highly likely that they do not perceive themselves as anonymous. Another potential explanation for the absence of this type of gendered language in the comments could be effective monitoring by DN.

The fact that the strongest predictor for a female article was the word *make* ("husband") would suggest that there still exists some degree of gendered language. Since "wife" did not appear among the top 40 male predictors it might be the case that more significance is attached to women's marital status than men's. However, the word "husband" only appeared eleven times across 3,246 articles. Similarly, female predictors like *medmänsklighet* ("compassion") and *känsla* ("feeling") (see Table 16 in Appendix 10.2) which are arguably communal attributes also constituted a small part of the

overall vocabulary.

In contrast to the evidence provided by Kittilson and Fridkin (2008), Table 5 and Table 6 jointly suggest that there is no clear gender division of political issues. Potentially, the distinction between "male" and "female" policy issues used by Kittilson and Fridkin (2008) is less relevant to the Swedish context during the time period under study. However, I would hesitate to declare that there are no gender divisions with regards to policy areas given that some of the strongest predictors for a female article made reference to specific policy areas.

Table 5 and Table 6 jointly suggest that professional titles occurred more often in articles about men. Since one of the strongest predictors for a male articles was the word "professor", which appeared 107 times in the gendered sample, one contributing factor may be the under-representation of women in other parts of society, such as academia.

The results in Table 5 and Table 6 also suggest that words related to family occurred more often in articles about women. While these words do not necessarily refer to a person's own family it likely reinforces the idea that women are, or should be, more concerned with issues related to the family. My empirical approach did not allow me to asses whether this observation is the result of a journalistic bias or whether women actively choose to work on issues related to family. A third possible explanation is that there may be institutional biases within political parties that encourage men and women to take different roles.

There are a few important limitation in the analysis presented above. First, I only had a limited sample of comments and it is likely that the lasso-logistic regression would have performed better if there was more data.

Second, I have not been able to reliably measure the number of people mentioned in each article. It may be the case that longer male articles were a result of a larger number of people being mentioned in those articles.

Third, traditional content analysis, used for example by Kittilson and Fridkin (2008), allows the researcher to understand what words are used in relation to a specific person. I have used a quantitative approach which allowed me to answer whether a word appeared in a given article, but not the context of the usage of the word.

Fourth, I have made an implicit assumption that there is a causal link between the person in the article and the vocabulary used in the article. This is not unreasonable, but it should be noted that the lasso-logistic regression is modelling the reverse relationship: given a set of words, what is the likelihood that the article is about a female? As a result, I have been able to show which words appeared disproportionally often in articles about women. However, the analysis above is not necessarily sufficient to assess the underlying causal relationship.

I have speculated that the observed gender gap is likely a reflection of real world disparities in the distribution of power between men and women. Assessing whether this is actually the case, I believe, would be a fruitful avenue for future research.

In addition to understanding the underlying cause of women's under-representation in political media, there are several outstanding descriptive questions. For example, if both men and women are mentioned in the same article, do they get an equal amount of space? Are there any gender patterns in the use of images?

Gendered language has many aspects, and the analysis presented here is not conclusive. Future research may inquire whether the tone (positive or negative) differs between articles about men and women, and whether there are other aspects of gendered language which I have not addressed here. One example may be further investigation into the hierarchy of naming (Messner et al., 1993): is one gender more often referred to by the first name rather than the last name?

9 Conclusions

All over the world women are largely absent from the media coverage of politics and government. If women are included they are often portrayed through a gendered lens which disproportionally emphasises personal aspects over professional aspects. This thesis has examined whether this phenomena exists in Swedish political reporting.

Using a sample of text data from Dagens Nyheter (DN), Sweden's largest daily newspaper, I have shown that there is indeed a difference between the amount of coverage that men and women receive. 2,508 articles mentioned only men, whereas 738 articles mentioned only women. I showed that articles featuring only women (who were not party leaders) were on average 8.7% shorter than articles featuring only men (who were not party leaders or prime minister). However, I have not found evidence that articles featuring women in party leadership positions were shorter. Similarly, while articles featuring women (who were not party leaders) on average received fewer comments, this was not the case for women in party leadership positions.

I have not found any significant correlation between the gender of the journalist and the amount of coverage received by men and women respectively; both in terms of number of articles, and article length.

To analyse the potential existence of gendered language I have used two approaches: topic analysis and lasso-logistic regression. The results from the topic analysis showed that professional titles occurred more frequently in articles about men: 77.5% of male articles mentioned at least one professional title and the corresponding share of female articles was 66.7%. In contrast, 10.9% of male articles mentioned at least one word related to family, whereas the corresponding share of female articles was 17.1%.

The results from the lasso-logistic regression provided weak evidence of the existence of gendered language in the article text. Words that might be considered indicative of its existence, such as *make* ("husband") and *medmänsklighet* ("compassion"), were rare; across 738 articles they appeared eleven and ten times respectively.

I have hypothesised that the gender disparities in media coverage are primarily a reflection of real world gender imbalances; both within politics, and other parts of society. One of the strongest predictors for a male article was the word "professor", suggesting that the experts featured in articles about politics are predominantly male.

Gendered language is a complex topic, its interpretation is highly context specific, and its subtleties provide insight into underlying societal inequities. Language is not only influenced by societal divisions, but language itself shapes our perceptions of reality: "Languages uses us as much as we use language" (Lakoff, 1973, p. 45).

References

Albrecht, J., Björklund, A., and Vroman, S. (2003). Is there a glass ceiling in Sweden? *Journal of Labor Economics*, 21(1):145–177.

Beaman, L., Chattopadhyay, R., Duflo, E., Pande, R., and Topalova, P. (2009). Powerful women: Does exposure reduce bias? *The Quarterly Journal of Economics*, 124(4):1497–1540.

Beaman, L., Duflo, E., Pande, R., and Topalova, P. (2012). Female leadership raises aspirations and educational attainment for girls: A policy experiment in India. *Science*, 335(6068):582–586.

Beltran, J., Gallego, A., Huidobro, A., Romero, E., and Padró, L. (2020). Male and female politicians on twitter: A machine learning approach.

Bem, S. L. and Bem, D. J. (1973). Does sex-biased job advertising "aid and abet" sex discrimination? *Journal of Applied Social Psychology*, 3(1):6–18.

Bertrand, M. (2018). Coase lecture–the glass ceiling. *Economica*, 85(338):205–231.

Bertrand, M., Black, S. E., Jensen, S., and Lleras-Muney, A. (2019). Breaking the glass ceiling? The effect of board quotas on female labour market outcomes in Norway. *The Review of Economic Studies*, 86(1):191–239.

Bettinger, E. P. and Long, B. T. (2005). Do faculty serve as role models? The impact of instructor gender on female students. *American Economic Review*, 95(2):152–157.

Bohnet, I. (2016). *What works*. Harvard University Press.

Bromander, T. (2012). *Politiska skandaler!: Behandlas kvinnor och män olika i massmedia?* PhD thesis, Linnaeus University Press.

Carlana, M. (2019). Implicit stereotypes: Evidence from teachers' gender bias. *The Quarterly Journal of Economics*, 134(3):1163–1224.

Carrell, S. E., Page, M. E., and West, J. E. (2010). Sex and science: How professor gender perpetuates the gender gap. *The Quarterly Journal of Economics*, 125(3):1101–1144.

Dagens Nyheter AB (2020a). *Audience*. Retrieved 17 May 2020, from https://annons.dn.se/en-se/audience.

Dagens Nyheter AB (2020b). *Om Dagens Nyheter*. Retrieved 17 May 2020, from https://www.dn.se/nyheter/om-dagens-nyheter.

Dagens Nyheter AB (2020c). *Svensk politik*. Retrieved 5 April 2020, from https://www.dn.se/om/svensk-politik.

Dagens Nyheter AB (2020d). *Så kan du kommentera artiklar på DN Debatt*. Retrieved 17 May 2020, from https://www.dn.se/arkiv/debatt/sa-kan-du-kommentera-artiklar-pa-dn-debatt.

De Cabo, R. M., Gimeno, R., Martínez, M., and López, L. (2014). Perpetuating gender inequality via the internet? An analysis of women's presence in Spanish online newspapers. *Sex Roles*, 70(1-2):57–71.

Dee, T. S. (2007). Teachers and the gender gaps in student achievement. *Journal of Human Resources*, 42(3):528–554.

Eagly, A. H. and Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, 109(3):573.

Edström, M. and Jacobsson, J. (2015). *Räkna med kvinnor Global Media Monitoring Project 2015: Nationell rapport Sverige.* Institutionen för journalistik, medier och kommunikation (JMG).

Esteve-Volart, B. and Bagues, M. (2012). Are women pawns in the political game? Evidence from elections to the Spanish Senate. *Journal of Public Economics*, 96(3-4):387–399.

European Institute for Gender Equality (2019). *Gender Equality Index 2019: Sweden.* Retrieved 17 May 2020, from https://eige.europa.eu/publications/gender-equality-index-2019-sweden.

Folke, O. and Rickne, J. (2020). All the single ladies: Job promotions and the durability of marriage. *American Economic Journal: Applied Economics*, 12(1):260–87.

Fox, R. L. and Lawless, J. L. (2004). Entering the arena?: Gender and the decision to run for office. *American Journal of Political Science*, 48(2):264–280.

Gallagher, M. (2005). *Who Makes the News?: Global Media Monitoring Project 2005.* World Association for Christian Communication.

Gaucher, D., Friesen, J., and Kay, A. C. (2011). Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology*, 101(1):109.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–74.

Global Media Monitoring Project (GMMP) (2015). *Who Makes the News?* World Association for Christian Communication (WACC).

Haraldsson, A. and Wängnerud, L. (2019). The effect of media sexism on women's political ambition: Evidence from a worldwide study. *Feminist Media Studies*, 19(4):525–541.

Hastie, T. and Qian, J. (2016). *An introduction to glmnet.*

Hsieh, C.-T., Hurst, E., Jones, C. I., and Klenow, P. J. (2019). The allocation of talent and us economic growth. *Econometrica*, 87(5):1439–1474.

Inglehart, R., Haerpfer, C., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., Lagos, M., Norris, P., Ponarin, E., Puranen, B., and et al. (eds.) (2014). *World Values Survey: Round six.* Retrieved 17 May 2020, from www.worldvaluessurvey.org/WVSDocumentationWV6.jsp.

Institute for Democracy and Electoral Assistance (IDEA) (2020). *Gender Quotas Database.* Retrieved 17 May 2020, from https://www.idea.int/data-tools/data/gender-quotas/country-view/261/35.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R.* Springer.

Jarlbro, G. (2006). *Medier, genus och makt*. Studentlitteratur.

Jönsson, A. M. (2005). Mångfalden i journalistkåren (The variety among journalists). *Göteborg: JMG, arbetsrapport*, 28.

Kahn, K. F. (1994). The distorted mirror: Press coverage of women candidates for statewide office. *The Journal of politics*, 56(1):154–173.

Kahn, K. F. and Goldenberg, E. N. (1991). Women candidates in the news: An examination of gender differences in us senate campaign coverage. *Public Opinion Quarterly*, 55(2):180–199.

Kittilson, M. C. and Fridkin, K. (2008). Gender, candidate portrayals and election campaigns: A comparative perspective. *Politics & Gender*, 4(3):371–392.

Koivula, N. (1999). Gender stereotyping in televised media sport coverage. *Sex Roles*, 41(7-8):589–604.

Kroon, Å. (2006). The gendered practice and role of pull quoting in political newspaper journalism. *News from the Interview Society. Göteborg: Nordicom*.

Kuhn, P., Shen, K., and Zhang, S. (2018). *Gender-targeted job ads in the recruitment process: Evidence from China*. Technical report, National Bureau of Economic Research.

Kurtulus, F. A. and Tomaskovic-Devey, D. (2012). Do female top managers help women to advance? A panel study using EEO-1 records. *The ANNALS of the American Academy of Political and Social Science*, 639(1):173–197.

Kyung, M., Gill, J., Ghosh, M., Casella, G., et al. (2010). Penalized regression, standard errors, and Bayesian Lassos. *Bayesian Analysis*, 5(2):369–411.

Lakoff, R. (1973). Language and woman's place. *Language in Society*, 2(1):45–79.

Lühiste, M. and Banducci, S. (2016). Invisible women? Comparing candidates' news coverage in Europe. *Politics & Gender*, 12(2):223–253.

Madera, J. M., Hebl, M. R., and Martin, R. C. (2009). Gender and letters of recommendation for academia: agentic and communal differences. *Journal of Applied Psychology*, 94(6):1591.

McGrath, K. (2014). Teaching sex, gender, transsexual, and transgender concepts. *Communication Teacher*, 28(2):96–101.

Messner, M. A., Duncan, M. C., and Jensen, K. (1993). Separating the men from the girls: The gendered language of televised sports. *Gender & Society*, 7(1):121–137.

Myndigheten för press, radio och TV (2019a). *Applying for a press and media subsidy*. Retrieved 17 May 2020, from https://www.mprt.se/en/applying-for-a-press-subsidy/.

Myndigheten för press, radio och TV (2019b). *Dagstidnigsförteckning 2019*. Retrieved 17 May 2020, from https://www.mprt.se/sv/blanketter–publikationer/publikationer/dagstidningsforteckning/.

Newman, N., Fletcher, R., Kalogeropoulos, A., and Nielsen, R. (2019). *Reuters institute digital news report 2019*, volume 2019. Reuters Institute for the Study of Journalism.

Reporters without borders (2020). *2020 World press freedom index*. Retrieved 17 May 2020, from https://rsf.org/en/ranking.

Statistics Sweden (2014). *På tal om kvinnor och män*. Stockholm: Statistics Sweden.

Statistikmyndigheten SCB (2020a). *Kvinnors inkomst närmar sig mäns – men långsamt*. Retrieved 17 May 2020, from https://www.scb.se/hitta-statistik/artiklar/2020/kvinnors-inkomst-narmar-sig-mans–men-langsamt/.

Statistikmyndigheten SCB (2020b). *Sveriges befolkning*. Retrieved 17 May 2020, from https://www.scb.se/hitta-statistik/sverige-i-siffror/manniskorna-i-sverige/sveriges-befolkning/.

Statistikmyndigheten SCB (2020c). *Utbildningsnivån i Sverige*. Retrieved 17 May 2020, from https://www.scb.se/hitta-statistik/sverige-i-siffror/utbildning-jobb-och-pengar/utbildningsnivan-i-sverige/.

Sveriges Riksdag (2020). *Elections to the Riksdag*. Retrieved 17 May 2020, from https://www.riksdagen.se/en/how-the-riksdag-works/democracy/elections-to-the-riksdag/.

The Economist (2020). *Iceland leads the way to women's equality in the workplace*. Retrieved 17 May 2020, from https://www.economist.com/graphic-detail/2020/03/04/iceland-leads-the-way-to-womens-equality-in-the-workplace.

The World Bank (2019). *Labor force, female (% of total labor force) - Sweden*. Retrieved 17 May 2020, from https://data.worldbank.org/indicator/SL.TLF.TOTL.FE.ZS?locations=SE.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

UN Women (2019). *Women in politics: 2019*. Retrieved 17 May 2020, from https://www.unwomen.org/en/digital-library/publications/2019/03/women-in-politics-2019-map.

UN Women (2020). *Women in politics: 2020*. Retrieved 17 May 2020, from https://www.unwomen.org/en/digital-library/publications/2020/03/women-in-politics-map-2020.

Vanden Broucke, S. and Baesens, B. (2018). *Practical Web scraping for data science*. Springer.

Wendt, M. (2011). *Landsfäder och småbarnsmammor: mediala gestaltningar av kön och politik*. Studentlitteratur.

West, C. and Zimmerman, D. H. (1987). Doing gender. *Gender & Society*, 1(2):125–151.

World Economic Forum (2019). *Global Gender Gap Report 2020*. ISBN-13: 978-2-940631-03-2.

Wu, A. H. (2017). Gender stereotyping in academia: Evidence from economics job market rumors forum. *Unpublished manuscript*.

Wu, A. H. (2018). Gendered language on the economics job market rumors forum. *AEA Papers and Proceedings*, 108:175–79.

10 Appendix

10.1 Figures

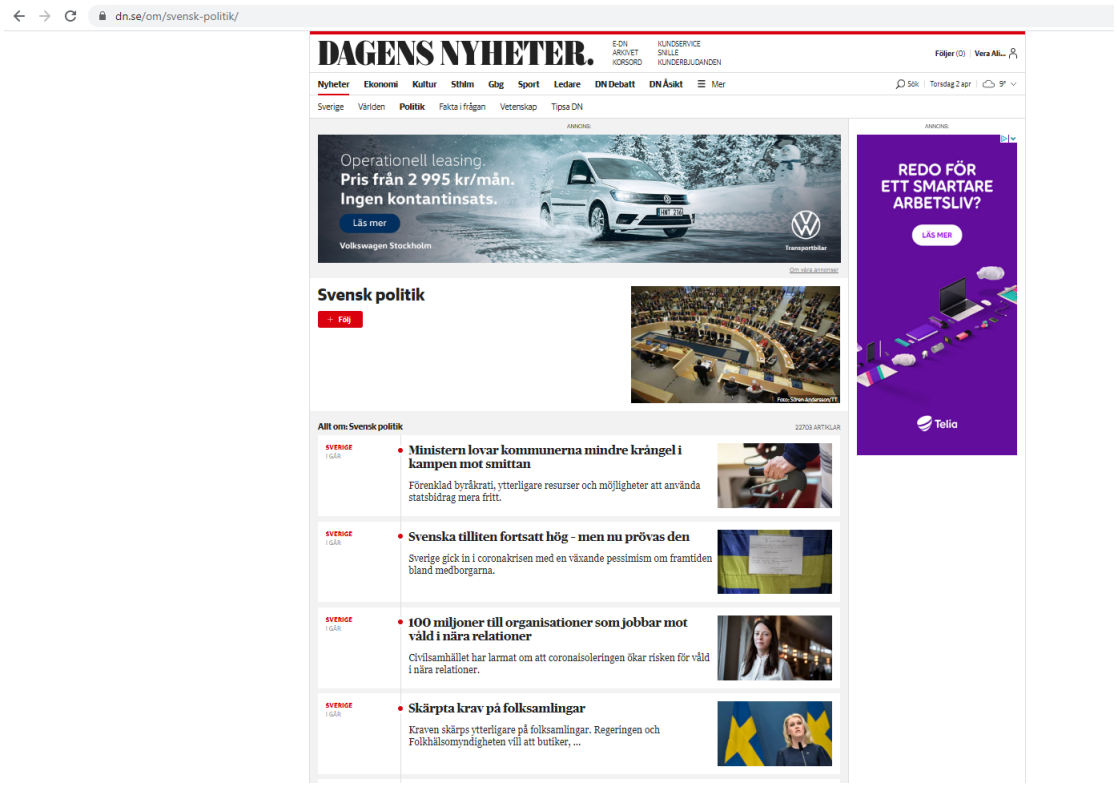
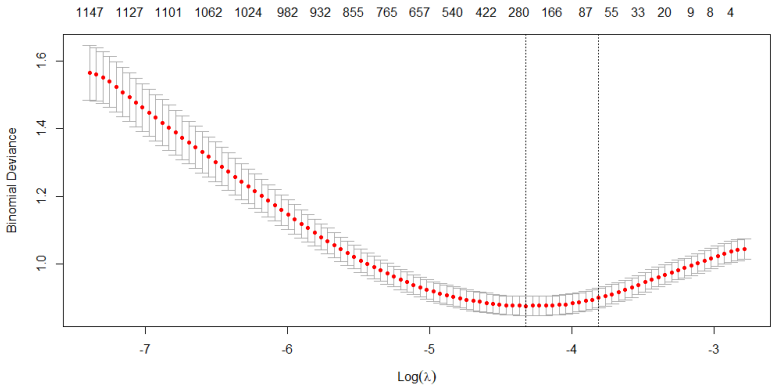
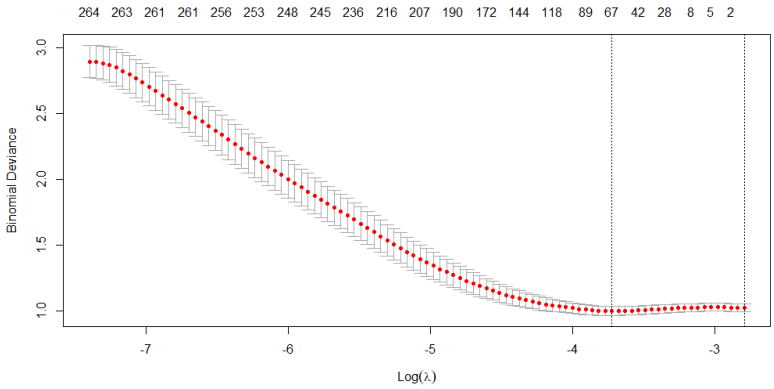


Figure 2: Snapshot of the politics section of DN



(a) Articles



(b) Comments

Figure 3: Number of predictors for different values of λ
Note: the number of predictors is provided at the top of each figure (ranging from 0 to 1147 in Figure 3a and from 0 to 264 in Figure 3b).

10.2 Tables

Party	Leader	Tenure
Social Democratic Party	Stefan Löfven	2012-
Moderate Party	Fredrik Reinfeldt	2003–2015
	Anna Kinberg Batra	2015-2017
	Ulf Kristersson	2017-
Sweden Democrats	Jimmie Åkesson	2005-
Green Party	Åsa Romson	2011-2016
	Gustav Fridolin	2011-2019
	Isabella Lövin	2016-
	Per Bolund	2019-
Centre Party	Annie Lööf	2011-
Left Party	Jonas Sjöstedt	2012-
Liberals	Jan Björklund	2007–2019
	Nyamko Sabuni	2019-
Christian Democrats	Göran Hägglund	2004-2015
	Ebba Busch	2015-

Table 11: Party leaders

	Female	Male
Pronouns	hon, henne, hennes, Hon, Henne, Hennes	han, honom, hans, Han, Honom, Hans
Names	Acko, Aida, Alexandra, Alice, Amanda, Amelia, Amineh, Andrea, Angela, Angelica, Angelika, Ann, Anna, Anna-Caren, Anna-Karin, Ann-Britt, Ann-Charlotte, Ann-Christin, Ann-Christine, Anne, Annelie, Annie, Annicka, Annika, Ann-Sofie, Aylin, Azadeh, Barbro, Beatrice, Betty, Birgitta, Boriana, Brita, Camilla, Carin, Carina, Carola, Caroline, Cassandra, Catharina, Cecilia, Cecilie, Charlotte, Christina, Ciczie, Clara, Corazza, Dana, Désirée, Ebba, Ellen, Elin, Elisabeth, Emilia, Emma, Eva, Fadime, Gita, Greta, Gulan, Gudrun, Gunilla, Gunvor, Hanna, Helén, Helena, Heléne, Helene, Helle, Hillevi, Ida, Ilona, Inga-Lill, Ingela, Ingrid, Irene, Isabella, Janine, Jennie, Jenny, Jessica, Jessika, Johanna, Josefin, Julia, Juno, Karin, Karolina, Katarina, Katja, Katrin, Kerstin, Kirsten, Kristina, Laila, Lauren, Lawen, Leila, Lena, Lina, Linda, Linnea, Lisa, Lorena, Lotta, Louise, Magdalena, Maj, Malin, Mara, Margareta, Margot, Maria, Marianne, Marie, Marie-Louise, Marit, Marlene, Marléne, Marta, Martina, Matilda, Mia, Mona, Monica, Monika, Märta, Nina, Nooshi, Noria, Nyamko, Paula, Pernilla, Pia, Rebecka, Rossana, Roza, Saila, Sandra, Sara, Sofia, Solveig, Soraya, Stina, Teres, Teresa, Tess, Tina, Ulla, Ulrika, Ursula, Vasiliki, Veronica, Veronika, Victoria, Yasmine, Ylva, Åsa	Abboud, Abdikani, Abdirahim, Adam, Adolf, Adnan, Aleksander, Alexander, Alf, Ali, Alireza, Allan, Amir, Anders, Andreas, Ardalan, Arin, Arman, Aron, Bengt, Benjamin, Bill, Birger, Björn, Bo, Boris, Carl, Carl-Oskar, Carl-Otto, Christer, Christian, Christofer, Christoffer, ClasGöran, Curt, Dacian, Dag, Dan, Daniel, David, Dawit, Denis, Dennis, Donald, Edward, Emil, Eric, Erik, Eskil, Folke, Fredrick, Fredrik, Gabriel, Gavin, Geerth, George, Gunnar, Gustaf, Gustav, Göran, Hampus, Hanif, Harald, Henrik, Hun, Håkan, Ibrahim, Ilmar, Ingemar, Isak, Jabar, Jacob, Jakob, Jakop, Jamal, Jan, Jasenko, Jens, Jimmie, Jimmy, Joakim, Joar, Joeri, Johan, John, Johnny, Jon, Jonas, Jonny, José, Josef, Jyrki, Jörgen, Kalle, Karl-Petter, Karsten, Kenneth, Kent, Kjell, Kjell-Arne, Klas, Lage, Larry, Lars, László, Leif, Linus, Lorentz, Ludvig, Magnus, Malcolm, Manuel, Marcello, Marcus, Mark, Markus, Martin, Matheus, Mats, Mathias, Mattias, Michael, Mikael, Momodou, Morgan, Neven, Niels, Nicklas, Nigel, Niklas, Ola, Olle, Olof, Oscar, Patrick, Patrik, Paul, Per, Per-Arne, Peter, Petter, Philip, Pyry, Pål, Rasmus, Richard, Rickard, Rikard, Robert, Robin, Roger, Roland, Rolf, Runar, Sebastian, Selahattin, Serkan, Soran, Staffan, Stefan, Stellan, Sten, Sultan, Svante, Sven-Olof, Sören, Thomas, Thoralf, Tobias, Tommy, Toivo, Tomas, Tony, Torbjörn, Torgny, Tuve, Ulf, Urban, Viktor, William, Örjan

Table 12: Article classifiers

	Female	Male	Organisation
Names	Agnes Laurell, Alexandra Carlsson Tenitskaja, Alexandra Urisman Otto, Amanda Dahl, Amanda Lindholm, Amanda Lindström, Amina Manzoor, Anette Nantell, Anna-Lena Laurén, Anna Gustafsson, Anna Kyringer, Anna Skoog, Annie Sääf, Annika Sohlander Cassel, Annika Ström Melin, Annika Wilhelmson, Annika Wilhelmson, Caroline Englund, Caroline Gyllenkrok, Cecilia Jacobsson, Elin Lindwall, Ewa Stenberg, Hanna Grosshög, Hanna Jakobson, Hanna Lilja, Ida Yttergren, Isabelle Nordström, Jannike Kihlberg, Jessica Ritzén, Johanna Sundbeck, Josefin Sköld, Josefine Stenersen, Karin Eriksson, Karin Grahn-Wetter, Katarina Lagerwall, Kristina Hedberg, Lina Lund, Lisa Röstlund, Lotta Härdelin, Malin Hansson, Maria Crofts, Maria Gunther, Maria Westholm, Marianne Björklund, Marijana Dragic, Marit Sundberg, Mathilda Ejefalk, Mia Holmgren, Mimi Billing, Pi Frisk, Pia Gripenberg, Raffaella Lindström, Ronja Mårtensson, Sandra Pandevski, Sandy Kalleny, Sanna Torén Björling, Thea Mossige-Norheim, Tina Zenou, Tindra Englund, Tove Nandorf, Ulrika By, Viviana Canoilas	Adam Darab, Adam Svensson, Alexander Kuronen, Alexander Mahmoud, Anders Bolling, Anders Boström, Anders Forsström, Anders Hansson, Andreas Lindberg, Anton Säll, Augustin Erba, Axel Björklund, Behrang Behdjou, Carl-Johan Kullving, Carl Cato, Caspar Opitz, Clas Svahn, Dan Lucas, Daniel Ågren, Daniel Costantini, David Ahlin, Edgar Mannheimer, Erik Ask, Erik de la Reguera, Erik Esbjörnsson, Erik Ohlsson, Fredrik Samuelson, Fredrik Tano, Hans Olsson, Hans Rosén, Hasse Eriksson, Hugo Ewald, Hugo Lindkvist, Ingmar Nevéus, Ivan Solander, Jack Werner, Jan Lewenhagen, Jens Kärrman, Jens Littorin, Jimmy Persson, Johan Esk, Johan Furusjö, Johan Schück, Johannes Ledel, Johar Bendjelloul, John Falkirk, Jonas Backlund, Jonas Desai, Josef Svenberg, Juan Flores, Kalle Holmberg, Karl Dalén, Kevin Chang, Kristoffer Örstadius, Kristoffer Törnmalm, Lars Näslund, Lasse Swärd, Lasse Wierup, Linus Larsson, Love Ahlstrand, Lukas Hansson, Magnus Hallgren, Måns Mosesson, Marcus Andersson, Martin Ezpeleta, Martin Gelin, Martin Gradén, Martin Jönsson, Mats J Larsson, Mattias Carlsson, Michael Winiarski, Mikael Bondesson, Mikael Delin, Mikael Holmström, Nathan Shachar, Nicklas Thegerström, Niklas Orrenius, Ossi Carp, Peter Letmark, Peter Loewe, Peter Wolodarski, Philip Teir, Robert Holender, Simon Frid, Simon Markusson, Staffan Kihlström, Stefan Lisinski, Sujay Dutt, Sverker Lenas, Torbjörn Petersson, Torbjörn Tenfält, Viktor Barth-Kron	DN, DN-TT, DN:s ledarredaktion, TT, TT-AFP, TT-DN, TT-Reuters

Table 13: Author classifiers

Topic	Swedish words (stems)	English translation
Professional titles	chef, finansminist, försvarsminist, gruppled, justitieminist, ledamot, ledamöt, minist, ordför, partiled, partisekreter, president, profess, riksdagsledamot, riksdagsledamöt, språkrör, statsminist, statsvet, talesperson, talman, utbildningsminist, utrikesminist	manager, minister for finance, minister of defence, parliamentary leader, minister of justice, member of parliament, members of parliament, minister, chairperson, party leader, party secretary, president, professor, member of parliament, members of parliament, spokesperson (party leader), prime minister, political scientist, spokesperson, speaker of the house, minister of education, minister of foreign affairs
Family words	barn, familj, föräldr	child/children , family, parents
Male issues	arbetsförmedl, arbetsgiv, arbetsmarknad, ekonomi, ekonomin, eu, fn, försvaret, försvarsmak, kärnkraft, militär, nato, skatt, skattesänkning	the public employment service, employer, labour market, economy, the economy, the european union, the united nations, the defence, the armed forces, nuclear power, military, the north atlantic treaty organization (nato), tax, tax reduction
Female issues	elev, friskol, jämstalld, miljö, pension, pensionär, sjukvård, skol, skolan, universitet, utsläpp, våldtäk, välfärd, vård	student, charter school, gender equality, environment, pension, pensioner, healthcare, school, school, university, emissions, rape, welfare, care

Table 14: Topic classifiers

Table 15: Results for number of articles (gendered sample)

	Dependent variable:	
	Female article	
	(1)	(2)
Election month	-0.074***	
	(0.020)	
Prime minister		-0.164***
		(0.020)
Party leader		-0.122***
		(0.017)
Constant	0.240***	0.322***
	(0.008)	(0.010)
Observations	3,246	3,246
R^2	0.004	0.070
Adjusted R^2	0.004	0.070
Residual Std. Error	0.418 (df = 3244)	0.404 (df = 3243)
F Statistic	14.122*** (df = 1; 3244)	122.625*** (df = 2; 3243)
Note:		*p<0.1; **p<0.05; ***p<0.01

Predictor (Swedish)	Translation (English)	Coefficient	n
make	husband	1.71	11
klimatminister	environment minister	1.65	11
centerledaren	the leader of the centre party	1.53	41
gymnasie	high school	1.51	10
jämställdhetsminister	minister for gender equality	1.41	34
arbetsmarknadsminister	minister of labour	1.27	36
socialförsäkringsminister	minister for social security	1.25	10
miljöminister	environment minister	1.15	26
migrationspolitisk	migration politics	1.08	12
religion	religion	0.99	10
informera	inform	0.97	19
handeln	the trade	0.86	12
kvinnors	women's	0.80	52
feministiskt	feminist	0.76	146
klimatpolitik	environmental politics	0.74	35
upptäcka	discover	0.74	16
röstkort	voting card	0.73	13
demokratiminister	minister of democracy	0.70	14
bensinskatt	petrol tax	0.69	11
jämtland	jämtland (province)	0.65	16
värmland	värmland (province)	0.63	13
enkät	survey	0.60	15
roligt	fun	0.59	20
kvinnan	the woman	0.58	23
arbetsgivaravgift	general payroll tax	0.56	11
medmänsklighet	compassion	0.56	10
finansminister	minister of finance	0.56	137
kopplas	coupled	0.53	21
beräknas	calculated	0.51	57
studio	studio	0.51	37
näringsdepartementet	ministry of enterprise and innovation	0.51	10
åriga	year old	0.49	20
önskat	wished	0.47	11
anpassat	adjusted	0.47	12
socialminister	minister for health and social affairs	0.45	48
medicinsk	medical	0.45	10
miljöministern	environment minister	0.43	14
massor	lots	0.43	12
känsla	feeling	0.43	25
avstår	abstain	0.43	21

Table 16: Strongest predictors for female article

Predictor (Swedish)	Translation (English)	Coefficient	n
(Intercept)	(Intercept)	-1.06	
statsminister	prime minister	-0.45	768
skillnader	differences	-0.38	62
sverigedemokraternas	the Sweden Democrats'	-0.37	516
heta	called, or hot	-0.32	18
professor	professor	-0.31	107
finanserna	finance	-0.22	35
statsministern	the prime minister	-0.22	385
förre	former	-0.21	83
vänsterledaren	the left party leader	-0.17	32
sverigedemokraterna	the Sweden Democrats	-0.16	1,275
ledaren	the leader	-0.15	721
hota	threaten	-0.14	17
utbildningsminister	minister of education	-0.14	88
flytta	move	-0.13	48
ger	gives	-0.13	442
vapen	weapon	-0.13	70
barth	barth (town or name)	-0.12	18
dåvarande	of that time	-0.11	63
riksdagens	parliament's	-0.11	406
opinionschef	opinion manager	-0.11	71
universitet	the university	-0.10	152
nye	new	-0.09	30
socialdemokraternas	the social democrats	-0.09	409
efter	after	-0.09	2,146
anklagelserna	the allegations	-0.09	24
höll	held	-0.08	127
talet	the speech	-0.08	365
demoskops	demoskop's (company)	-0.08	16
varit	been	-0.08	1,120
mellan	between	-0.08	1,284
uppgifterna	information	-0.07	101
mannen	the man	-0.07	109
sina	their	-0.07	1,053
december	december	-0.06	202
händelsen	the event	-0.06	61
premiärminister	prime minister	-0.06	33
samlar	collect, or gather	-0.05	48
ingen	no one	-0.05	658
ett	one	-0.05	9,212

Table 17: Strongest predictors for male article

Predictor (Swedish)	Translation (English)	Coefficient	n
kvinna	woman	1.03	11
behov	needs	0.73	12
skatt	tax	0.68	19
medborgare	citizen	0.65	23
sådant	such	0.62	18
senare	later	0.58	10
innebär	means	0.50	19
landsting	county	0.48	11
barn	child	0.46	21
skatter	taxes	0.45	13
någon	someone	0.43	96
staten	the state	0.40	19
artikeln	the article	0.33	22
vid	by	0.33	48
lärare	teacher	0.26	13
utbildning	education	0.25	14
flyktingar	refugees	0.23	10
samtidigt	simultaneously	0.21	25
större	bigger	0.18	28
bra	good	0.18	117
finns	exist	0.18	135
sedan	since	0.17	57
förstår	understand	0.16	28
fortsätter	continue	0.14	11
vilka	which	0.13	32
räcker	suffice	0.11	12
visst	sure	0.09	10
svt	svt (national public television broadcaster)	0.09	14
varit	been	0.05	55
leva	live	0.05	16
minst	least	0.05	15
kräva	demand	0.03	10
gjort	done	0.03	29

Table 18: Strongest predictors for female (comment text)

Predictor (Swedish)	Translation (English)	Coefficient	n
(Intercept)	(Intercept)	-1.24	
valet	the election	-0.55	40
partierna	the parties	-0.46	32
emot	against	-0.43	25
naturligtvis	naturally	-0.40	16
förstå	understand	-0.38	21
liten	small	-0.35	14
haft	had	-0.30	22
resurser	resources	-0.27	15
partiet	the party	-0.21	42
ingen	no one	-0.16	91
fast	but, or firm	-0.15	16
form	shape	-0.14	12
ett	one	-0.13	444
ska	will	-0.12	203
dom	they	-0.12	69
mindre	less	-0.11	26
moderaterna	the moderate party	-0.10	34
och	and	-0.10	1,480
mot	against	-0.09	87
områden	areas	-0.08	16
men	but	-0.08	194
liksom	like	-0.06	10
när	when	-0.06	184
sig	itself	-0.05	258
byta	change	-0.05	10
håller	holding	-0.04	20
regering	government	-0.03	82
stället	the place	-0.02	13
brott	crime	-0.02	30
intressant	interesting	-0.02	32
röster	votes	-0.02	18
kommer	coming	-0.01	154
partier	parties	-0.01	60
ser	looking	-0.00	48

Table 19: Strongest predictors for male (comment text)