Measuring Gender Bias in News Images

Sen Jia
Department of Computer Science
University of Bristol
Bristol, United Kingdom
jason.jia@bris.ac.uk

Thomas Lansdall-Welfare
Department of Computer Science
University of Bristol
Bristol, United Kingdom
thomas.lansdall-welfare@bris.ac.uk

Nello Cristianini
Department of Computer Science
University of Bristol
Bristol, United Kingdom
nello.cristianini@bris.ac.uk

ABSTRACT

Analysing the representation of gender in news media has a long history within the fields of journalism, media and communication. Typically this can be performed by measuring how often people of each gender are mentioned within the textual content of news articles. In this paper, we adopt a different approach, classifying the faces in images of news articles into their respective gender. We present a study on 885,573 news articles gathered from the web, covering a period of four months between 19th October 2014 and 19th January 2015 from 882 news outlets. Findings show that gender bias differs by topic, with Fashion and the Arts showing the least bias. Comparisons of gender bias by outlet suggest that tabloid-style news outlets may be less gender-biased than broadsheet-style ones, supporting previous results from textual content analysis of news articles.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications—*Computer Vision, Text Processing*; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Object recognition*

General Terms

Experimentation, Measurement

Keywords

Gender Bias, News Analysis, Image Classification

1. INTRODUCTION

It is well known that news media influences public opinion [15, 16, 17], and considerable attention has been paid by social scientists to the detection of any biases in news reporting. One of the most studied ones is bias in gender representation: how many males and females are represented in the news narrative, and how does this representation change by topic, by outlet or by country.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742007.

Traditionally, both text and image analysis have been performed by hand [3, 14, 18], but in recent times the detection of gender bias in text has been automated [1, 6], making it possible to analyse hundreds of thousands of news articles, from hundreds of different news outlets, over a period of months. This revealed a strong pattern of male over-representation, which is however strongly dependent on topic: Entertainment being less skewed than Politics, in the representation of genders. Interestingly, those patterns were correlated with analogous differences in income inequality between the genders.

Now we provide a large scale study of gender bias in the images used by the news media. This involves using face detection and classification tools, which have now reached very high levels of performance even for images taken under uncontrolled conditions, such as those that are found in the news. We discover again a pattern of male over-representation in the media, but with interesting variations by topic and type of newspaper.

Previous studies on text from news media found that in a collection of British and Irish news from the Global Media Monitoring Project [14], males are three times more visible than females [18] and that within particular outlets (CNN, FOX, New York Times) males are more dominant in their number of mentions [3]. Our own studies on gender bias in news media text found that gender bias correlates with topic and choice of news outlet [1, 6], with males being far more common in Sports, Politics and Business, while Fashion and Arts are less gender bias.

This paper will briefly describe the infrastructure for the detection and analysis of articles in news articles, as well as their topic. We will not discuss in detail the technology for image classification (which is described elsewhere, and beyond the scope of this paper), but we will provide experimental measurements on the performance of our face gender classifier, benchmarking it on the standard dataset for face gender recognition: LFW, specifically created to represent images of people's faces taken in uncontrolled conditions [8].

Our study involves: 460,493 images from 885,573 articles, published by 882 news outlets, all in English language, gathered in the period from 19th October 2014 to 19th January 2015.

2. METHODOLOGY

We gathered 885,573 news articles from the web between 19th October 2014 and 19th January 2015 using our modular architecture for news media analysis [5, 7]. Previously, this system has been successfully used for several media anal-

Label	Description	Count
No URL	No image URL was present in the news article header meta information.	377,269
Broken URL	No article image could be retrieved from the URL.	47,811
Logo	The same article image has appeared on at least 10 different articles.	133,715
No Face	The article image does not contain a face.	178,652
Face	The article image does contain a face.	148,126

Table 1: Description of the news article image labels and their counts.

ysis studies in both news and social media, ranging from analysing public mood from social media [11], large-scale analysis of topic, style and gender bias in news text content [6], and detecting the framing of scientific coverage in the media [12].

News articles were collected from the main RSS feed of 882 news outlets, filtering out any articles which were not written in English. As part of the system, we automatically classified each news article into 15 different generic news categories, such as "Crime" or "Science" using Support Vector Machines (SVM) [4] trained for high precision on the New York Times [19] and Reuters corpora [13] along with online linear perceptron models [4] trained on news media within our modular architecture. A total of 521,108 articles were assigned to at least one topic category, with topic categories being non-exclusive.

For each news article in our corpus, we retrieved the image URL specified in the header meta information of the web page the article was retrieved from, additionally recording each time an article did not have this property. If a URL was associated with 10 or more articles, we designated it as a "Logo" image. Following this, we took each URL extracted from the news articles, and attempted to download the image, performing face recognition on the image using the Viola-Jones algorithm [10] in OpenCV [2]. If a face was present, we further performed gender classification on the largest face in the image, assigning a label of either "Male" or "Female".

For gender classification, we extracted multi-scale LBP features from each face image, before applying our gender classification algorithm [9], based upon an online C-Pegasos [20], attaining an accuracy of over 95% on the Labeled Faces in the Wild (LFW) dataset [8].

We assessed the performance of our gender classifier under realistic conditions by applying our classifier to the LFW dataset, rescaled to mimic the sizes that we can expect to find in our news article images.

After this procedure, we could label each news article with one of six possible labels, as detailed in Table 1, along with the number of news article falling into each category. Those news articles that were labelled "Face" were also further labelled as "Male" or "Female", depending on the face gender recognised by the gender classification algorithm.

3. FINDINGS

In our first set of experiments, we compared the proportion of male faces to female faces, denoted as the male/female ratio, in the news article images we collected, and looked at how this male/female ratio differed for each of the 15 topic categories.

In the second set of experiments, we compare the male/female ratio of the news articles by their publishing outlet, showing the extent of the gender bias towards males present in each outlet.

3.1 Gender Bias by Topic

We wanted to investigate how the number of men and women featured in the images associated with the news differs across topics. As a first test, to ensure that results are comparable across topics, we show the percentage of images assigned each image label per topic in Figure 1. We can see that for each of the 15 topics, approximately 30–50% of the articles do not have an associated image, either because there was no associated image, or the image URL returned no image. Another 10–20% of the articles only have logo images, leaving 30–55% of the articles with a non-logo image.

For those articles which have a face image, we find the proportion of faces assigned to each gender and compute the male/female ratio, as shown in Figure 2. We can immediately see that in all topics except Fashion, there is a higher number of male faces than female faces, with Arts and Entertainment having the next lowest ratios of males to females. This is perhaps unsurprising, with other studies [1, 6] finding similar results for these topics. The topic categories of Petroleum, Politics and Markets contain the largest gender bias, with almost four or more male faces to every female face that is displayed.

While we measure the male/female ratio of faces in news media images, there is an additional real-world bias related to each topic which should be kept in mind. The male/female ratio of people working in related fields to each of these topics is not necessarily equal, and therefore can influence how often a male or female face is likely to be pictured in the news.

3.2 Gender Bias in Outlets

We also wanted to investigate how the gender bias of male to female faces in news images differed by the news outlet that published the images. Again, to ensure we could make fair comparisons across outlets, we show the image label breakdown for 20 well known outlets selected from the corpus in Figure 4. We can see that for some outlets (The Times, New York Times, The Telegraph, Chicago Tribune) nearly every single news article did not contain a URL to an image, suggesting that these outlets simply do not use the header meta information as a way to associate news images with their article counterparts. We can also see that some outlets (Fox, Guardian, Seattle Times) are very likely to use an logo as the image for the article, instead of an image that represents the article content. For these reasons, we do not include these outlets in our following analysis of gender bias by outlet.

For the remaining 12 outlets, we can see their male/female ratio in Figure 3. We can see that the tabloid-style outlets (Daily Mail, Daily News, Metro) tend to have a lower gen-

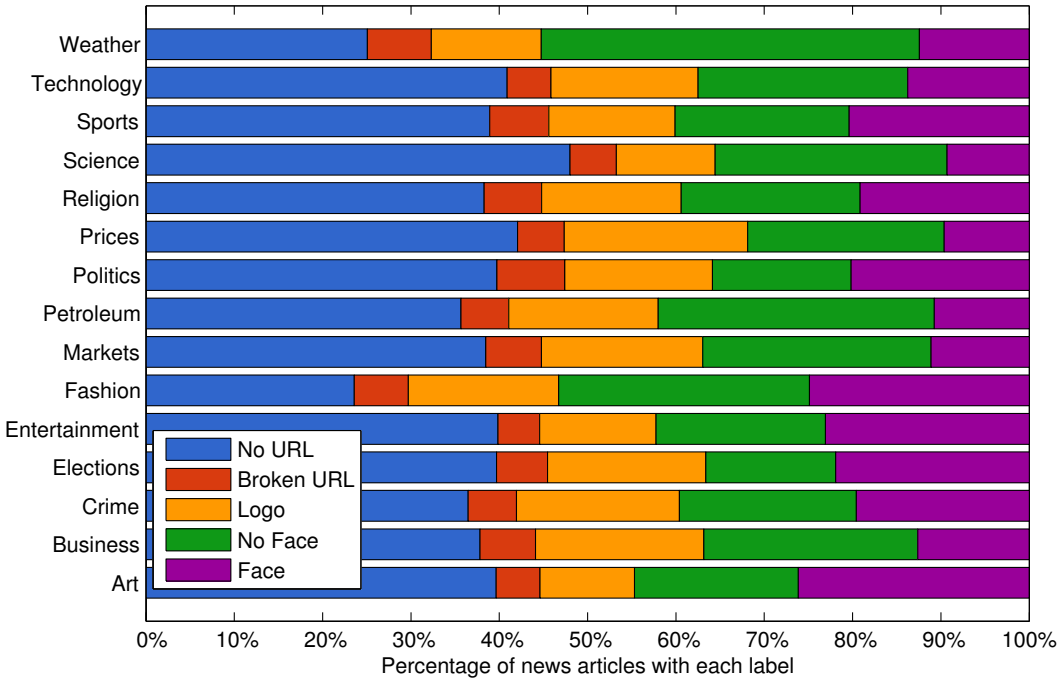

Figure 1: Percentage of each type of image label assigned to the topics.

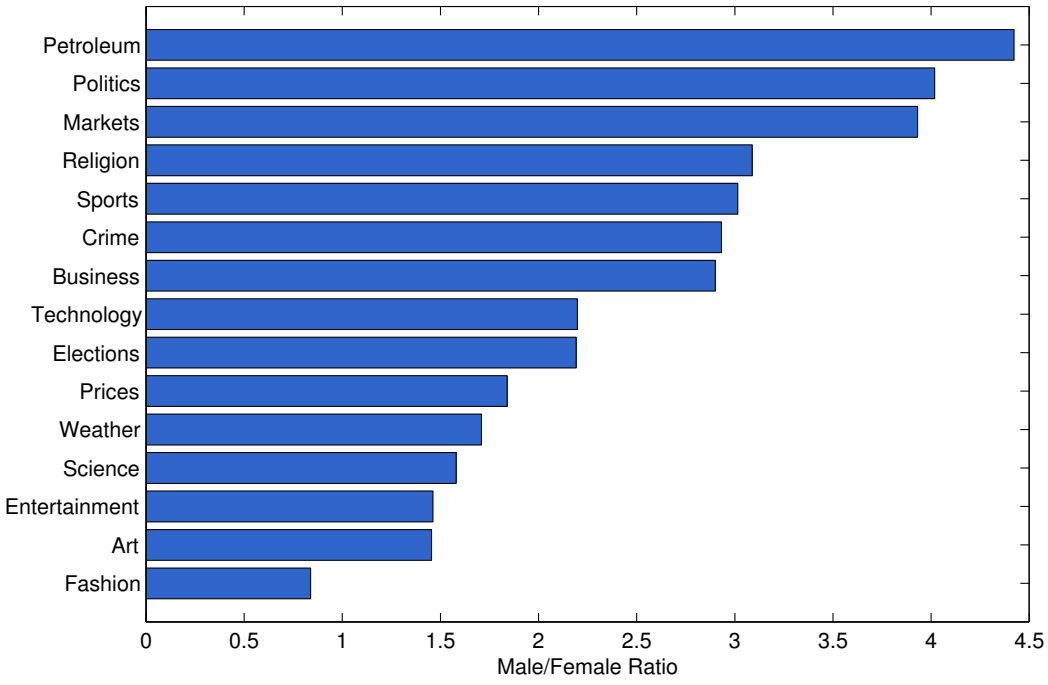

Figure 2: Male/Female ratio for each topic.

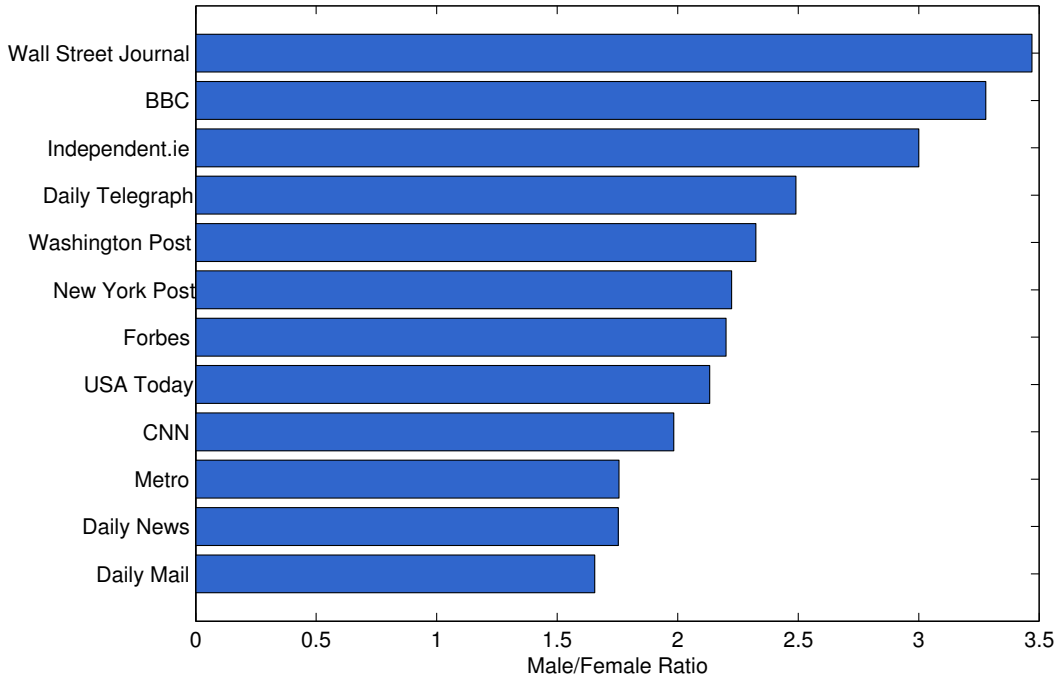

Figure 3: Male/Female ratio for each outlet.

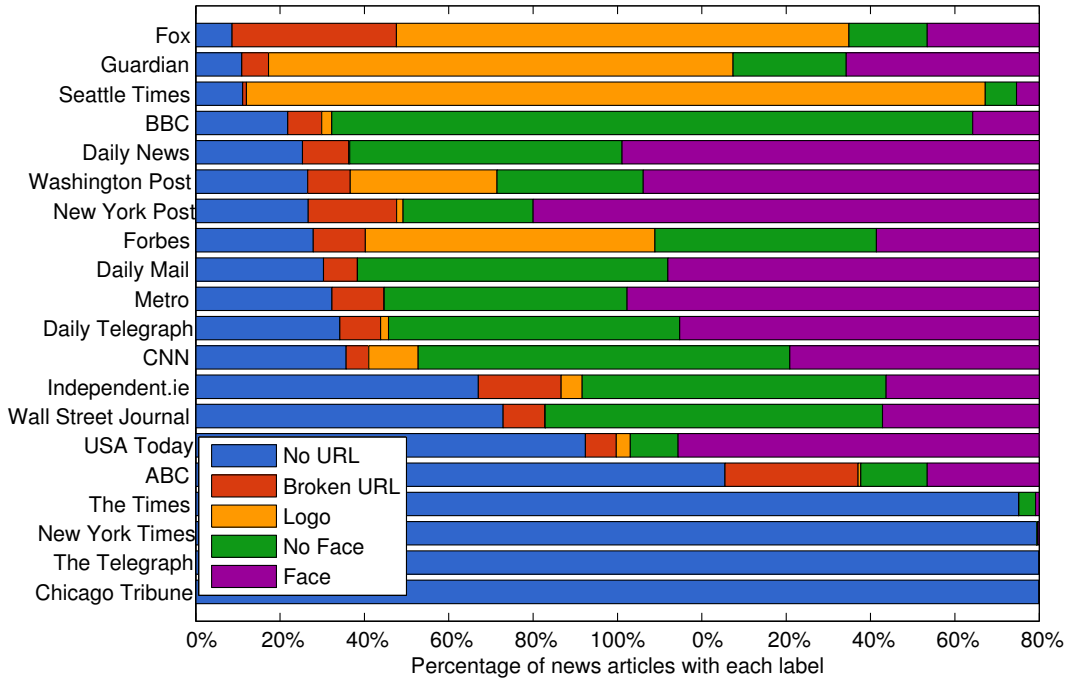

Figure 4: Percentage of each type of image label assigned to each outlet.

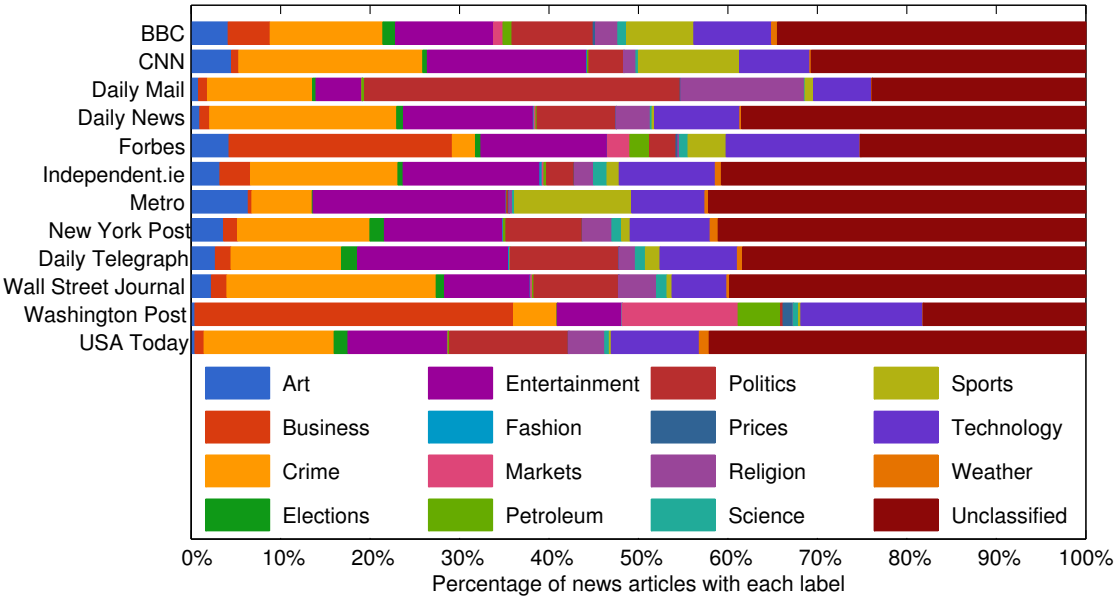

Figure 5: Distribution of topic classifications for each outlet.

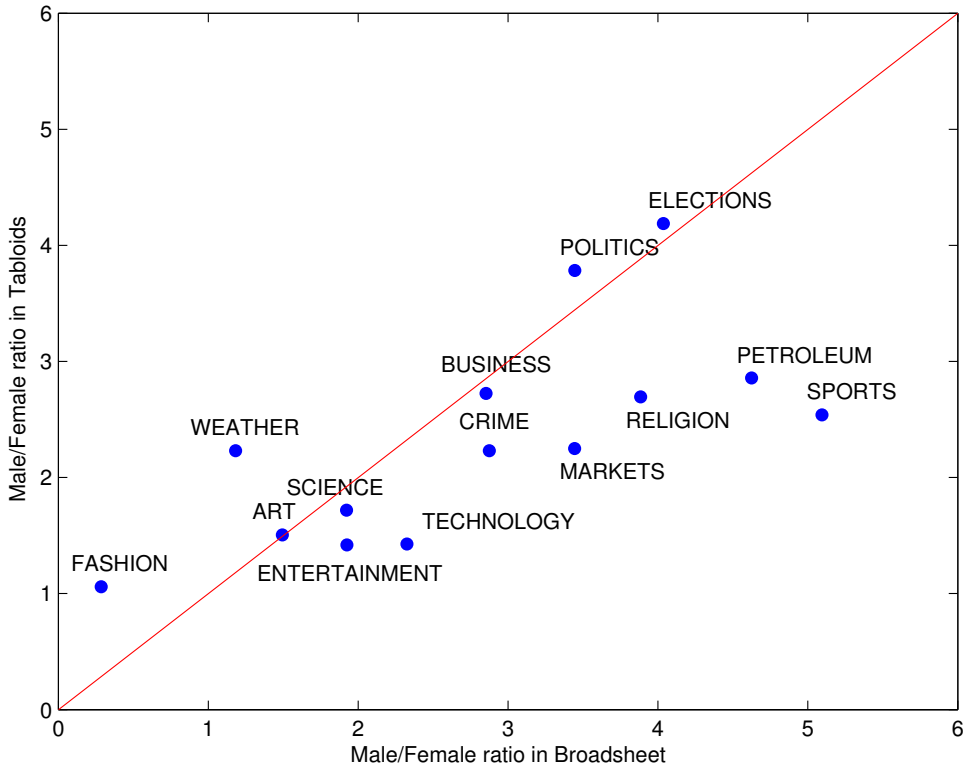

Figure 6: Male/Female ratio for outlets grouped by type per topic.

der bias, while broadsheet-style outlets tend to have a larger gender bias. This difference could be down to the different topic distributions of the given outlets, where tabloid-style outlets focus more heavily on topics which have a lower gender bias, while the broadsheet-style outlets focus on higher gender bias topics such as Politics, Religion, Crime and Business. To further investigate this, Figure 5 shows the topic distributions for each of the 12 outlets. The Daily Mail has a large proportion of Politics articles by topic classification (a high gender bias topic), but overall as an outlet has the lowest gender bias. This suggests that the Daily Mail does indeed feature more images of female faces in its news coverage, taking into account the topic distribution of the outlet. Another example of this is the Washington Post. where a large percentage of its coverage is devoted to Business and Markets; topics where the average face gender ratio is roughly 3–4 male faces to each female face. Despite this, the Washington Post has an average gender bias ratio of 2.3 male faces to each female face.

We additionally grouped the outlets into tabloid-style (Daily Mail, Daily News, Metro, New York Post) and broadsheet-style (BBC, CNN, Forbes, Independent.ie, Daily Telegraph, Wall Street Journal, Washington Post) outlets to compare the gender bias between these two types of outlets[1], with the results shown in Figure 6. Topics that fall below the red line have a higher male/female ratio in broadsheet-style outlets, while topics above the red line have a higher male/female ratio in tabloid-style outlets[2]. We can see that tabloid-style outlets have a higher gender bias in Fashion and Weather topics than their broadsheet-style counterparts, while broadsheet-style outlets have a comparatively higher gender bias in Markets, Petroleum, Religion, Sports and Technology.

4. CONCLUSIONS

The automation of news content analysis is a very promising direction in the social sciences, but it has so far been mostly limited to text [1, 6]. The possibility of analysing the content of images in a very large scale paves the way to a completely new set of studies, gender bias being only one of the possibilities. It is now also possible to investigate the interplay between text and images, and how people are portrayed and represented in different countries, or different topics.

This study is intended as a feasibility study, but also it is the first study of this type to our knowledge.

5. ACKNOWLEDGMENTS

Sen Jia is supported by the EU Project ThinkBIG, while Thomas Lansdall-Welfare and Nello Cristianini are supported by the EU Projects Complacs and ThinkBIG.

6. REFERENCES

[1] Omar Ali, Ilias N Flaounas, Tijl De Bie, Nick Mosdell, Justin Lewis, and Nello Cristianini. Automating News Content Analysis: An Application to Gender Bias and Readability. 2010.

[2] Gary Bradski. The OpenCV Library. *Doctor Dobbs Journal*, 25(11):120–126, 2000.

[1]Excluding USA Today as it does not clearly fall into either type.

[2]The 'Prices' topic featured no female faces in tabloid-style outlets so a male/female ratio could not be computed, for broadsheet-style outlets it had a ratio of 1.5.

[3] Cindy Burke and Sharon R Mazzarella. "A Slightly New Shade of Lipstick": Gendered Mediation in Internet News Stories. *Women's Studies in Communication*, 31(3):395–418, 2008.

[4] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, 2000.

[5] I. Flaounas, O. Ali, M. Turchi, T. Snowsill, F. Nicart, T. De Bie, and N. Cristianini. NOAM: News Outlets Analysis and Monitoring System. In *SIGMOD 2011*, pages 1275–1278. ACM, 2011.

[6] Ilias Flaounas, Omar Ali, Thomas Lansdall-Welfare, Tijl De Bie, Nick Mosdell, Justin Lewis, and Nello Cristianini. Research Methods in the Age of Digital Journalism: Massive-Scale Automated Analysis of News Content - Topics, Style and Gender. *Digital Journalism*, 1(1):102–116, 2013.

[7] Ilias Flaounas, Thomas Lansdall-Welfare, Panagiota Antonakaki, and Nello Cristianini. The Anatomy of a Modular System for Media Content Analysis. *CoRR*, abs/1402.6208, 2014.

[8] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[9] Sen Jia and Nello Christianini. Learning to Classify Gender from Four Million Images. *Pattern Recognition Letters*, 2015, DOI:10.1016/j.patrec.2015.02.006.

[10] Michael Jones and Paul Viola. Fast Multi-View Face Detection. *Mitsubishi Electric Research Lab TR-20003-96*, 3:14, 2003.

[11] Thomas Lansdall-Welfare, Vasileios Lampos, and Nello Cristianini. Effects of the Recession on Public Mood in the UK. In *Proceedings of the 21st International Conference Companion on World Wide Web*, WWW '12 Companion, pages 1221–1226. ACM, 2012.

[12] Thomas Lansdall-Welfare, Saatviga Sudhahar, Giuseppe A. Veltri, and Nello Cristianini. On the Coverage of Science in the Media: A Big Data Study on the Impact of the Fukushima Disaster. In *Proceedings of the 2014 IEEE International Conference on Big Data*, 2014.

[13] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. RCV1: A New Benchmark Collection for Text Categorization Research. *The Journal of Machine Learning Research*, 5:361–397, 2004.

[14] Sarah Macharia, Dermot O'Connor, and Lilian Ndangam. *Who Makes the News?: Global Media Monitoring Project 2010*. World Association for Christian Communication, 2010.

[15] Maxwell E McCombs and Donald L Shaw. The Agenda-Setting Function of Mass Media. *Public opinion quarterly*, 36(2):176–187, 1972.

[16] Maxwell E McCombs and Donald L Shaw. The Evolution of Agenda-Setting Research: Twenty-Five Years in the Marketplace of Ideas. *Journal of communication*, 43(2):58–67, 1993.

[17] Amy Reynolds and Maxwell E McCombs. News Influence on our Pictures of the World. *Media effects: Advances in theory and research*, 10:1–18, 2002.

[18] Karen Ross and Cynthia Carter. Women and News: A long and Winding Road. *Media, Culture & Society*, 33(8):1148–1165, 2011.

[19] Evan Sandhaus. The New York Times Annotated Corpus. *Linguistic Data Consortium, Philadelphia*, 6(12), 2008.

[20] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal Estimated Sub-Gradient Solver for SVM. *Mathematical programming*, 127(1):3–30, 2011.