JMLR: Workshop and Conference Proceedings 11 (2010) 36–43 Workshop on Applications of Pattern Analysis

Automating News Content Analysis: An Application to Gender Bias and Readability

Omar Ali OMAR.ALI@BRISTOL.AC.UK
Ilias Flaounas ILIAS.FLAOUNAS@BRISTOL.AC.UK
Tijl De Bie TIJL.DEBIE@BRISTOL.AC.UK
Intelligent Systems Laboratory, University of Bristol, UK

Nick Mosdell MOSDELLN1@CARDIFF.AC.UK
Justin Lewis LEWISJ2@CARDIFF.AC.UK
Cardiff School of Journalism, Media and Cultural Studies, Cardiff University, UK

Nello Cristianini NELLO.CRISTIANINI@BRISTOL.AC.UK
Intelligent Systems Laboratory, University of Bristol, UK

Editors: Tom Diethe, Nello Cristianini, John Shawe-Taylor

Abstract

In this article we present an application of text-analysis technologies to support social science research, in particular the analysis of patterns in news content. We describe a system that gathers and annotates large volumes of textual data in order to extract patterns and trends. We have examined 3.5 million news articles and show that their topic is related to the gender bias and readability of their content. This study is intended to illustrate how pattern analysis technology can be deployed to automate tasks commonly performed by humans in the social sciences, in order to enable large scale studies that would otherwise be impossible.

1. Introduction

The analysis of news content is an important part of modern social sciences, and is often aimed at disclosing subtle biases in the way news is reported. The presence and portrayal of gender in the news media has a long history within media and cultural studies, often involving complex judgements of stereotypes and language as well as the relative incidence of male and female sources and actors (Carter et al., 1998). The majority of these investigations are conducted by hand, and involve selecting small samples of news coverage and collecting information relevant to the study in a process known as 'coding.' This process requires a high level of attention to detail and must also be repeated independently in order to minimise human error and bias.

In this paper we describe a system that incorporates text-analysis technologies for the automation of some of these tasks, enabling us to extract patterns from news media coverage on a very large scale. We have used automated techniques to gather over 3.5 million on-line news articles and have extracted information from them, such as gender references and ease of readability.

The data we present here is illustrative of what can be achieved using automatic coding technology. Our findings suggest, first of all, a strong gender bias in the set of people covered

in the news. We found that references to men outnumber references to women by three to one, which suggests that little has changed since earlier studies of television (Gerbner et al., 1986). Secondly, this gender bias is shown to differ significantly in articles of different topics. We observe, for example, that articles about sports or business are among the most gender-biased, while articles in entertainment are the least gender-biased.

Thirdly, we measured readability of text and observed significant differences in readability of news articles across topics. The most readable topics are entertainment and sports, while the least readable are business and science.

Our fourth finding is the observation of a heavy-tailed distribution in the frequency of references to people, showing that most are concentrated on a small number of people, while the remainder are spread across a large number of people.

Previous work on gender bias was carried out by Len-Ríos et al. (2006), examining gender bias across seven topics. Similarly, Dalecki et al. (2009) investigated the readability of news articles. Len-Ríos et al. also found the most male bias in sports articles, with the least in entertainment. We note that they based their study on three weeks of articles from two newspapers; a tiny sample by comparison, but illustrative of the constraints of traditional coding techniques.

In contrast, our system can automate much of this process, enabling us to use a sample that would simply be beyond the reach of a traditional content analysis—examining 1,121 media outlets over a 14 month period. We believe that this approach presents considerable opportunities for interaction in the field of automatic pattern analysis, especially due to the availability of massive textual data sets.

2. Experiments

The conducted experiments were based on our system outlined in Figure 1. It collects news items from news outlets that have an on-line presence and offer their content using the Really Simple Syndication (RSS) format. This provides an easy means for machines to automatically collect the content of a web-site. We monitored 1,121 different outlets using a 'spider', which checks their feeds every five hours. RSS feeds only advertise the title and a short description of the article's content, along with a link to the main article. Thus, we deployed a 'scraper' to extract the full content of every article. This allowed us to collect around 15,000 articles per day. We manually tagged all RSS feeds with topic tags that are based on the section of the web-site from which they originate. These tags are automatically propagated to articles contained within a feed.

We examined a period of 14 months, between 1st January 2009 and 28th February 2010, during which we collected 3,529,125 English-language articles across six topics: 'top stories' (1,839,457 articles), 'entertainment' (746,581), 'science' (570,924), 'politics' (376,511), 'business' (239,285) and 'sport' (40,782)—note that some articles have mutliple tags. The 'top stories' tag is assigned to articles from the main page of media outlets. We deployed modules to extract information from these articles and, in this study, we consider named entity recognition (NER) with gender detection and the readability of articles.

Named Entity Recognition and Gender Labelling In this study we focus on references to people, extracted using the Named Entity Recognition (NER) tool contained in the open source suite, GATE (Cunningham et al., 2002). This tool can determine the

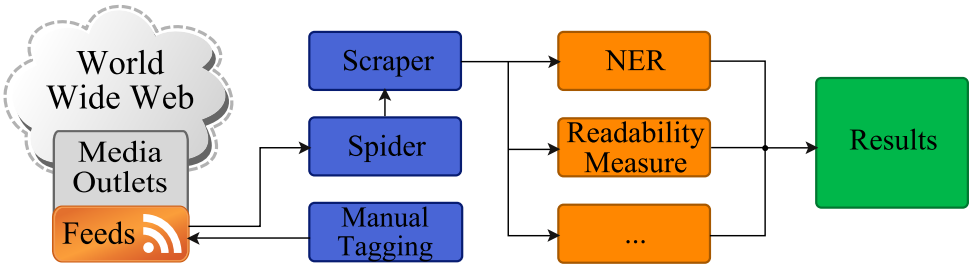

Figure 1: Our system architecture.

gender of a person based on various linguistic clues found in the surrounding text, but this assigment is noisy as it is based on unstructured text. Errors are possible as the result of rare occasions where a reference to a male entity is incorrectly linked to some female-indicator like the pronoun 'she'. In order to exercise caution, we designed our system to exploit the fact that every reference has been observed multiple times in our corpus, and we assign a 'Male' or 'Female' gender label only when the NER tool recommends it for at least 70% of sightings. When there is insufficient evidence, we consider a reference to be ambiguous and assign the label, 'Unsure'.

We evaluated the accuracy of this method by comparing it to Freebase, which is a structured on-line database.[1] From it we collected 38,480 database entries for people who could be matched exactly to the most popular references in our database. These Freebase entries provided us with gender labels that we use as the ground-truth for evaluation: 30,569 references were correctly labelled as male or female, 7,285 were labelled as 'Unsure' and only 626 cases were labelled incorrectly. We checked to ensure that we did not introduce bias when our module labelled references as 'Unsure'; in this case the percentage of references that are actually 'Female' is 16.2%. Similarly, across all 38,480 labels of our ground-truth, we find that 'Female' labels account for 17.7% of references. This indicates that we are not biased when applying the 'Unsure' label to references.

Measurement of Readability The ease of comprehension of an article was measured using the Flesch Reading Ease Test (Flesch, 1948). This empirical test is based on the linguistic properties of the input text, such as the average number of words per sentence and the average number of syllables per word. The resulting readability scores range from 0 to 100 and can be interpreted as follows: scores 90–100 correspond to texts easily understandable by an average 11-year-old student, scores 60–70 are easily understandable by 13-to-15 year old students and 0–30 are less readable and best understood by university graduates. We are aware of the limitations of this measure, which takes little account of the cultural aspects of language—it is, nonetheless, indicative of readability as well as providing an example of a suitable framework for automated coding.

1. Freebase, `http://www.freebase.com`.

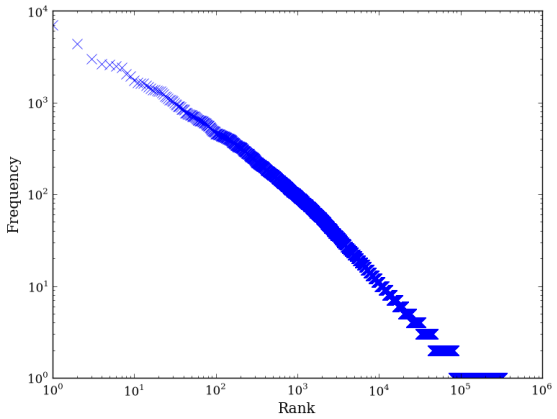

Figure 2: References to people seen between 1st January 2009 and 28th February 2010 exhibit a heavy-tailed distribution.

3. Results

Heavy-tailed Distribution of References Figure 2 shows that references to people exhibit a heavy-tailed distribution. This is compatible with a rich-gets-richer dynamic (Clauset et al., 2009), where popular people are more likely to attract news coverage, and is also consistent with findings from previous research into the gender of the 'primary definers' or most powerful sources within news media (Carter et al., 1998).

Gender Bias Results Our system examined articles for each month of the study period, counting the 1,000 most-frequent references to people in each topic of the study. It then counted the gender labels of these references, resulting in counts of 'Male', 'Female' and 'Unsure' labels, for each topic in each month.

Figure 3 shows the 'Male' and 'Female' reference counts among the top 1,000 ranked by frequency in each topic. We show the sample mean over all months of the study. 'Male' and 'Female' counts do not sum up to 1,000 as the remaining references are those that were labelled 'Unsure'.

In articles tagged as 'Top Stories' we find that 'Male' references make up 49.0% of the top 1,000, compared to only 18.6% for 'Female'. A similar bias is also found in 'Science' articles. This increases in 'Business' and 'Politics' articles and is most evident in 'Sport' articles, where we find 60.0% 'Male' references and only 6.7% 'Female'. On the other hand, we find 'Entertainment' articles to be less biased, with 51.2% 'Male' and 29.3% 'Female' references in the top 1,000.

As an additional test, we collected two months of data from 'Female Sports' feeds. These feeds contain articles that focus specifically on women's sports and consist of 807 articles, collected during January and February 2010. As we might expect, we found a significantly larger proportion of 'Female' references, but even in this category—where we might expect a strong bias towards women—nearly half the references are to males.

Gender Bias in the Distribution of Wealth In order to chart the extent to which media coverage of various topics reflects gender inequities in the broader world we compared

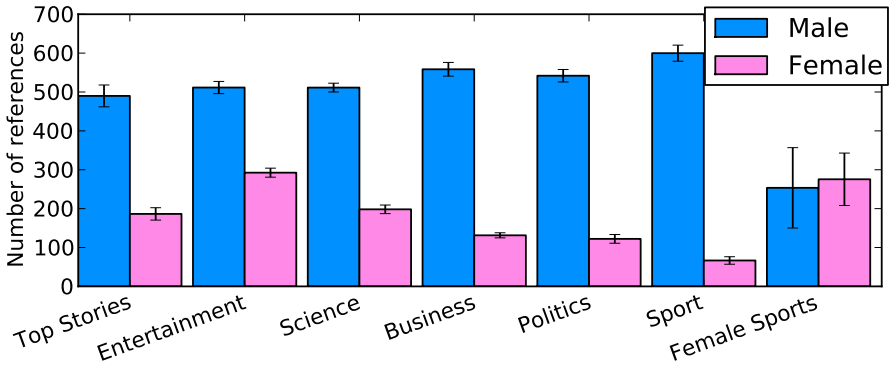

Figure 3: 'Male' and 'Female' reference counts for a selection of topic tags, averaged over all months of the study and sorted in order of 'Female' reference count. Error bars represent 95% confidence intervals of the sample mean.

Rich List	List Length	Male Count	Female Count	Female %
Celebrity	100	67	33	33.0
All	100	88	10	10.2
U.S. Politicians	16	15	1	6.25
Athletes	50	50	0	0.00

Table 1: Frequency of appearance of males and females in a selection of 'rich lists.' We examine the 100 richest celebrities, 100 richest people in the world, 16 richest U.S. politicians and 50 richest athletes.

our results to the distribution of gender among the wealthiest people in a selection of topic areas. Table 1 shows gender counts for three of our topic areas, collected from the following 'rich lists': 100 richest celebrities, 16 richest U.S. Politicians and 50 highest-paid athletes. We also include the 100 richest people in the world for comparison. We see no females in the richest athletes and only one amongst the richest U.S. politicians. Of the 100 richest people in the world, 10 are female and only in 'entertainment' do we see a less heavy male bias, with 33 females in the top 100.

This suggests that differences in news coverage reflect broader inequities amongst various elites. In this sense, coverage reflects the fact that (at the top end) sport is the most male dominated field, followed by 'politics,' with 'entertainment' being the least male-dominated field.

Readability Bias Results We measured readability of articles on 1,000 randomly sampled articles from selected topics of interest. We added a further test, using the BBC show CBBC-Newsround, which is a current affairs programme written specifically for children. As we would expect Newsround to demonstrate a higher readability rating than news targeted at adults, this provides a suitable test of the readability criteria.

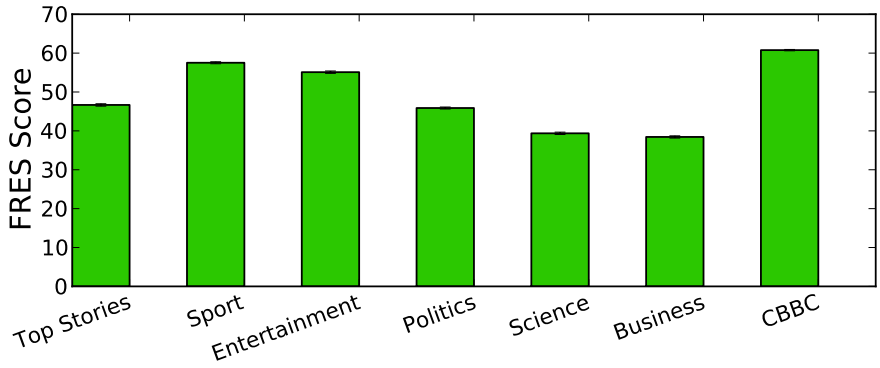

Figure 4: FRES scores on topics of interest. Error bars represent 95% confidence intervals of the sample mean.

Figure 4 shows the mean FRES score and error bars represent 95% confidence intervals. We found that 'Sport' articles were the most readable, followed by 'Entertainment,' 'Politics,' 'Science' and 'Business.' Articles from CBBC-Newsround were, as expected, highly readable.

Overall, we found a degree of gender inflection to the readability of topics, with 'Sport' stories the most readable, followed by 'Entertainment', which is also the least male dominated. This is consistent with previous studies of readability, using much smaller samples, that found sports and news stories featuring women to be far easier to read than those concerning international news (Stempel, 1981).

4. Conclusions

Gender bias is highly evident in media content, as it is in other parts of society, such as income distribution. Our contribution has been to measure this on a dataset of millions of articles, extracting and relating information on gender and readability.

While this study shows the feasibility of automated large scale analysis, and perhaps points in the direction of Computational Social Sciences, we should also discuss the limitations of this kind of approach. Firstly, it is currently limited to 'shallow semantics,' which offer more information than plain string-matching methods, but may miss nuances in the meaning of articles. Also, it is perhaps less accurate than human coders since a software tool is more likely to misunderstand something like the gender of a person than a human coder, who is aware of how cultural codes might override linguistic cues. On the other hand, this approach compensates for these drawbacks by enabling researchers to access vast datasets, and hence to apply the law of large numbers and statistical error correction. Furthermore, even if it is less accurate than human coding, certain experiments would not be possible if they were not automated.

While the rigidity of any automated system has its disadvantages, it is also less subject to forms of interpretative bias and measurement error than human forms of coding, which depend on the experimenter formulating a precise set of questions, so establishing a rigid

framework of analysis. Automatic coding makes it possible for patterns to emerge from the data, rather than simply in response to a coding frame. We will of course, still rely on human interpretation to recognise the significance of those patterns, but it does not depend upon our ability to anticipate which frameworks are dominant and which are not.

We have demonstrated the application of large scale data analysis methods to the domain of news content analysis. Much more can be done by automated means in this domain (Godbole et al., 2007; Pouliquen et al., 2004; Flaounas et al., 2010). Ultimately, we should be able to chart the flow of information in the news media environment, to see who is setting the agenda, how, and to what effect.

Acknowledgments

O. Ali is supported by an EPSRC Doctoral Training Grant; I. Flaounas is supported by the A. S. Onassis Public Benefit Foundation; N. Cristianini is supported by a Royal Society Wolfson Merit Award. Members of the Intelligent Systems Laboratory are supported by the 'Pascal2' Network of Excellence.

References

C. Carter, G. Branston, and S. Hall. *News, Gender and Power*. Routledge, 1998.

A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51:661–703, 2009.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proc. of the 40th Anniversary Meeting of the Association for Computational Linguistics*, pages 168–175, 2002.

L. Dalecki, D.L. Lasorsa, and S.C. Lewis. The news readability problem. *Journalism Practice*, 3(1):1–12, 2009.

I. Flaounas, N. Fyson, and N. Cristianini. Predicting relations in news-media content among EU countries. *2nd IAPR International Workshop on Cognitive Information Processing*, pages 269–274, 2010.

R. Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233, 1948.

G. Gerbner, L. Gross, M. Morgan, and N. Signorielli. Living with television: The dynamics of the cultivation process. In *Perspectives on media effects*, pages 17–40, 1986.

N. Godbole, M. Srinivasaiah, and S. Skiena. Large-scale sentiment analysis for news and blogs. *International Conference on Weblogs and Social Media*, 2007.

M.E. Len-Ríos, S. Rodgers, E. Thorson, and D. Yoon. Representation of women in news and photos: Comparing content to perceptions. *Journal of Communication*, 55(1):152–168, 2006.

B. Pouliquen, R. Steinberger, C. Ignat, E. Käsper, and I. Temnikova. Multilingual and cross-lingual news topic tracking. In *Proceedings of the 20th international conference on Computational Linguistics*, page 959. Association for Computational Linguistics, 2004.

G. H. Stempel. Readability of six kinds of content in newspapers. *Newspaper Research Journal*, 3(1):32–37, 1981.