Automating News Content Analysis
An Application to Gender Bias and Readability

Omar Ali, Ilias Flaounas, Tijl De Bie, Nello Cristianini,
Nick Mosdell, Justin Lewis.
Overview

• Examine topics in the media.
• Examples of research by social scientists.
• How we can automate such work?
• Demonstrate biases by topic:
  • Gender.
  • Readability.
Readability:
FRES = 51.8
References to:
Males = 6
Females = 2

Readability:
FRES = 66.9
References to:
Males = 2
Females = 2
Results from the Social Sciences

• The readability (FRES) of news articles is higher for fabricated stories (Dalecki, Lasorsa, Lewis, 2009).

• Gender bias in news articles varies across topics (Len-Rios et al., 2005):
  • Entertainment is the least male-biased.
  • Front Page.
  • Business.
  • Sport displays the most male-bias.
How can we automate this?
How can we automate this?

Brittany Murphy’s widower found dead - Simon Monjack has died just
Rachel Stevens is expecting! - Rachel Stevens has delighted her fans
Russell and Rossy to reunite - Jonathan Ross is reportedly ending his
Katy Perry tops Maxim Hot 100 list - Katy Perry might be about to tie the
Kylie, shy and retiring? Pahl - Kylie Minogue might be one of the most
Jude and Sienna: it’s official! - Jude Law and Sienna Miller have been
Sandra Bullock files for divorce, adopts baby - Sandra Bullock has filed
Jack Tweed cleared of rape charges - Jack Tweed, the widower of Jack
Mark Owen reunites with wife - Mark Owen recently admitted to having
Brian apologises for Twitter rant - It’s been a tough couple of years for
Diana Vickers on life after X Factor - Our music editor James Hurley in
Kim Kardashian dating Cristiano Ronaldo? - A couple of weeks ago,
Kerry Katona on her ‘car crash’ life - Kerry Katona has finally opened
Katie Price’s death fears - In an interview with OK! magazine, Katie Price
Kylie for Fashion Targets Breast Cancer - New pictures for the campaign
Cheryl’s voicemail rape threat drama - Detectives have launched an in
Has Isla Fisher become Mrs Borat? - They’ve been engaged for a while!
Sandra Bullock hit by Oscars curse? - Sandra Bullock has pulled out of
Katie to launch kids make-up range? - Katie Price is no stranger to controversy
Cheryl and Ashley finally meet - After weeks of being apart, Cheryl and
Kerry Katona signs Peter Andre’s manager - It’s not been an easy co
How can we automate this?
Title: Red Bull can be beaten – McLaren

Description: Championship leaders Red Bull can be caught this season, according to McLaren managing director Jonathan Neale.

URL: http://news.bbc.co.uk/sport1/hi/

Content: Red Bull have been in scintillating form this campaign, claiming pole in all six races and winning three. McLaren and the other teams have been left playing catch-up but Neale said: "They can be beaten. "Red Bull have a have a quick car and they have the upper hand but I am confident that won't be the case by the middle of the season." Red Bull lead both the drivers' and constructors' standings with Australian Mark Webber…
How can we automate this?

The system monitored:

- 6 topics
- 1,121 outlets
- 14 months
- 3.5 million English-language articles
How can we automate this?
Flesch Reading Ease Score

\[
206.876 - 1.015 \left( \frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left( \frac{\text{total syllables}}{\text{total words}} \right)
\]

Scores:

90.0–100.0: easily understandable by an average 11-year-old student

60.0–70.0: easily understandable by 13- to 15-year-old students

0.0–30.0: best understood by university graduates
## Gender Labelling

<table>
<thead>
<tr>
<th>Name</th>
<th>Gender</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barack Obama</td>
<td>Male</td>
<td>11,998</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>380</td>
</tr>
<tr>
<td>Cheryl Cole</td>
<td>Male</td>
<td>132</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>8,520</td>
</tr>
<tr>
<td>Hillary Clinton</td>
<td>Male</td>
<td>3,529</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>5,493</td>
</tr>
</tbody>
</table>
How can we automate this?

Spider

Scrapper

Articles

Gender Assignment

Named Entity Recognition

Readability (FRES)

Results
Summary

• Social scientists study the media.
• Online media is easily accessible to machines.
• We have automated collection and analysis of news.
• Demonstrated biases by topic:
  • Gender.
  • Readability.
**Evaluation against Freebase**

<table>
<thead>
<tr>
<th>Freebase Label</th>
<th>Our Label</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>Female</td>
<td>Unsure</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>25,142</td>
<td>411</td>
<td>6,108</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>215</td>
<td>5,427</td>
<td>1,177</td>
<td></td>
</tr>
</tbody>
</table>
Links