What did Big Data find when it analyzed 150 years of British history?

January 9 2017

Credit: CC0 Public Domain

What could be learnt about the world if you could read the news from over 100 local newspapers for a period of 150 years? This is what a team

of Artificial Intelligence (AI) researchers from the University of Bristol have done, together with a social scientist and a historian, who had access to 150 years of British regional newspapers.

The patterns that emerged from the automated analysis of 35 million articles ranged from the detection of major events, to the subtle variations in gender bias across the decades. The study has investigated transitions such as the uptake of new technologies and even new political ideas, in a new way that is more like genomic studies than traditional historical investigation.

The team of academics, led by Professor Nello Cristianini, collaborated closely with the company findmypast, who is digitising historical newspapers from the British Library as part of their British Newspaper Archive project.

The main focus of the study was to establish if major historical and cultural changes could be detected from the subtle statistical footprints left in the collective content of local newspapers. How many women were mentioned? In which year did electricity start being mentioned more than steam? Crucially, this work goes well beyond counting words, and deploys AI methods to identify people and their gender, or locations and their position on the map.

The landmark study, part of the University of Bristol's ThinkBIG project, collected a huge amount of regional newspapers from the UK, including geographical and time-based information that is not available in other textual data such as books. Over 35 million articles and 28.6 billion words, from the British Library's newspaper collections, representing 14 per cent of all British regional outlets from 1800 to 1950, were used for the study.

Nello Cristianini, Professor of Artificial Intelligence, from the

Department of Engineering Mathematics, said: "The key aim of the study was to demonstrate an approach to understanding continuity and change in history, based on the distant reading of a vast body of news, which complements what is traditionally done by historians.

"The research team showed that changes and continuities detected in newspaper content can reflect culture, biases in representation or actual real-world events. More detailed studies on the same data will be performed."

Simple content analysis allowed the researchers to detect specific key events like wars, epidemics, coronations or gatherings with high accuracy, while the use of more refined techniques from AI enabled the research team to move beyond counting words by detecting references to named entities, such as individuals, companies and locations.

Some of the results were to be expected, and acted as a rational check for the approach, while other outcomes were not so obvious at the start of the analysis.The researchers found in the areas of values, beliefs and UK politics that in the 19th century Gladstone was much more newsworthy than Disraeli; until the 1930's Liberals were mentioned more than Conservatives, and that reference to British identity took off in the 20th century.

In the subjects of technology and economy, the research team tracked the steady decline of steam and the rise of electricity, with a crossing point of 1898; trains overtook horses in popularity in 1902; and the four largest peaks for 'panic' corresponded with negative market movements linked to banking crises in 1826, 1847, 1857 and 1866.

The researchers have shown in the subjects of social change and popular culture that the Suffragette movement fell within a delimited time interval 1906 to 1918; 'actors', 'singers' and 'dancers' began to increase in

the 1890s, rising significantly from then on, while references to 'politicians', by contrast, gradually declined from the early 20th century; and that 'football' was more prominent than 'cricket' from 1909.

Replicating a previous study done on book content, the researchers then moved on to link famous people in the news to their profession, finding that politicians and writers are most likely to achieve notoriety within their lifetimes, while scientists and mathematicians are less likely to achieve fame but decline less sharply.

More importantly, the researchers found that males are systematically more present than females during the entire period studied, but there is a slow increase of the presence of women after 1900, although it is difficult to attribute this to a single factor at the time. Interestingly, the amount of gender bias in the news over the period of investigation is not very different from current levels.

Dr Tom Lansdall-Welfare, Research Associate in Machine Learning in the Department of Computer Science, who led the computational part of the study, said: "We have demonstrated that computational approaches can establish meaningful relationships between a given signal in large-scale textual corpora and verifiable historical moments.

"However, what cannot be automated is the understanding of the implications of these findings for people, and that will always be the realm of the humanities and social sciences, and never that of machines."

The researchers believe that these data-driven approaches can complement the traditional method of close reading in detecting trends of continuity and change in historical corpora.

'Content analysis of 150 years of British periodicals' by Lansdall-Welfare et al is published in *PNAS*.

More information: Content analysis of 150 years of British periodicals, *PNAS*, www.pnas.org/cgi/doi/10.1073/pnas.1606380114

Provided by University of Bristol