The University of Manchester

Q-Step

A step-change in quantitative social science skills

Funded by the Nuffield Foundation, ESRC and HEFCE

Gendered stereotypes of race: a Big Data analysis - Data Fellowship at Manchester China Institute

Ihesinachi Oyouwa Oko-Jaja | Department of Politics/Department of Crimnology

Overview of the Data Fellowship

Our project was inspired by BLM and #StopAsianHate. Drawing on the concepts of gendered stereotypes of race and stereotype content model literature and the semi-automated dictionary-based methodology of Nicolas, Bai and Fiske (2020) to determine whether people were using physicality stereotypes against Black and Asian men in the US.

- Our aim was to test the following hypotheses:
- HA: There will be a greater proportion of low physicality words used in comments related to Asian men than comments related to Black men.
- HA: There will be a greater proportion of high physicality words used in comments related to Black men than comments related to Asian men.

Data Analysis

We created high and low physicality stereotype dictionaries using the Wordnet R package. We used the RedditExtractoR package to scrape 6682 comments from 8 subreddits relating to Asian and Black communities, masculinities and issues of racism more generally. We then created two corpora: one for comments about Asian men and one for comments about Black men.

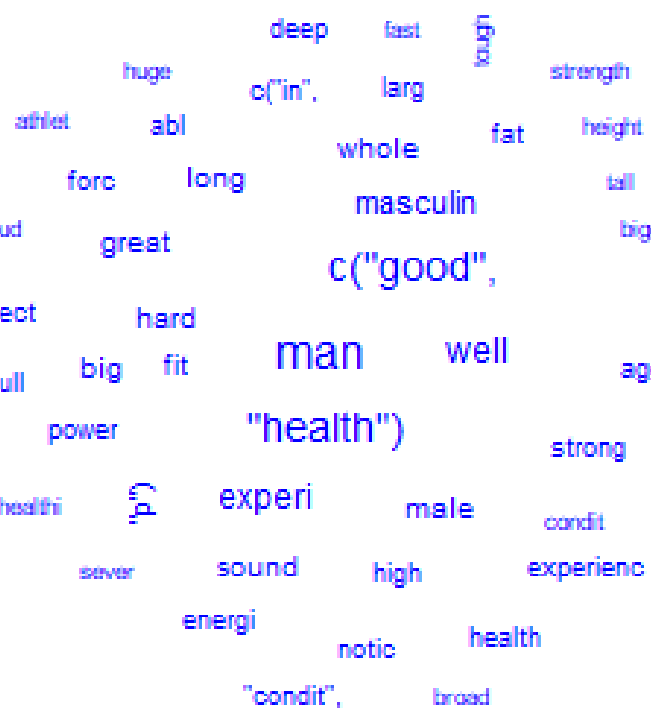

Figure 1: wordcloud of high physicality words in comments about African American men

After removing irrelevant words, punctuation and spaces in the comments, we used the Natural Language Processing model UDpipe to identify and isolate the top 1000 adjectives and nouns in each corpus. Finally, we used our dictionaries to identify the percentages of the top words in each corpus which were in our stereotype dictionaries.

Figure 2: wordcloud of top adjectives in comments in Asian men's subreddits

Findings

- There was a stronger association of low physicality words with Asian men than Black men (0.255% compared to 0.194%), supporting our hypotheses
- There was a slightly weaker association of high physicality words with Black men than Asian men (2.50% compared to 2.65%), contradicting our hypotheses
- Our hypotheses have only partially been supported by our analysis, partially because of limitations with the method such as difficulty of data cleaning, social desirability bias on social media, and unsuitability of data for inferential statistics

• Key Skills Learnt

Introduction to **corpus linguistics** and related techniques:

- Basic **natural language processing**
- New types of **data visualisation** in R - word clusters, dendrograms
- Semi-automated data scraping of Reddit in RStudio
- Cleaning, tokenisation and dictionary analysis in RStudio

Semi-automated creation of dictionaries (SADCAT, via SADCAT R package developed by Nicolas et al.

Dictionary analysis of textual data

Conversion of R data.frames into .csv files that can be opened in Microsoft Excel

Contact: **ihesinachi.oyouwaoko-jaja@student.manchester.ac.uk**