Media Bias Towards African-americans Before and After the Charlottesville Rally

Leschke, Julia C.; Schwemmer, Carsten

Erstveröffentlichung / Primary Publication
Konferenzbeitrag / conference paper

Empfohlene Zitierung / Suggested Citation:
Leschke, J. C., & Schwemmer, C. (2019). Media Bias Towards African-americans Before and After the Charlottesville Rally. In *Proceedings of the Weizenbaum Conference 2019 "Challenges of Digital Inequality - Digital Education, Digital Work, Digital Life"* (pp. 1-10). Berlin https://doi.org/10.34669/wi.cp/2.25

PROCEEDINGS OF THE WEIZENBAUM CONFERENCE 2019
CHALLENGES OF DIGITAL INEQUALITY
DIGITAL EDUCATION | DIGITAL WORK | DIGITAL LIFE

Media Bias towards African-Americans before and after the Charlottesville Rally

Julia C. Leschke
London School of Economics
and Political Science
London, England
J.Leschke@lse.ac.uk

Carsten Schwemmer
University of Bamberg
Bamberg, Germany
carsten.schwemmer@uni-bamberg.de

Abstract

African-Americans are still experiencing racial discrimination rooted in structural bias in US American society. Research has shown that this behaviour can be reduced if individuals are made conscious of their bias, but little is known about these mechanisms on a societal level. Envisaging the white-supremacist Charlottesville rally in 2017 as an event that rendered American society conscious of its racism, we scrutinise whether racial bias in the digital media has changed, comparing levels of pre- and post-Charlottesville bias. We fit word embedding models to a broad sample of largely US media and quantify bias by calculating cosine similarities between terms for black or white actors and positive or negative character traits. We find no differences in positive character traits after Charlottesville. However, African-Americans are associated substantially less with negative character traits post-Charlottesville, while white actors are semantically closer to negative traits.

Keywords

Media bias; Ethnic studies; Automated text analysis; Word embeddings; Charlottesville

DOI: 10.34669/wi.cp/2.25

1 Introduction

The African-American population is still confronted with racial discrimination, which originates from a negative structural bias of American society towards black people. A recent and extreme example of the discriminatory behaviour, with which African-Americans are confronted, is the high number of fatal shootings of unarmed black men by white police officers across the United States. Digital media, online news and blogs play a central role in the persistent phenomenon of racial discrimination, as they serve as a primary source of important information on current events but also, more generally, inform and shape the attitude and worldview held by the population. Due to its crucial role in opinion formation and updating, the implicit (as well as explicit) positive or negative bias towards minority groups spread by online media can reinforce biases in individuals, which can lead to discriminatory behaviour. In this sense, biases spread by media sources across the political spectrum of broadsheets, tabloids and blogs can be regarded as a proxy for public bias. In this study, we provide evidence in support of the idea that biases can be found across a broad spectrum of news sources and that these biases are likely to shift over time. We particularly focus on online media as a proxy for public opinion as well as the public's strength and direction of bias. Related to this issue we also ask how the biases that persist in digital societies emerge from the individual level.

Several studies have shown that making people conscious of their racial bias can pave the way to a significant reduction in discriminatory behaviour (Devine et al. 2012, Amodio, Devine, and Harmon-Jones 2007). Yet, can an intervention that reduces bias in individuals under laboratory conditions also work on digital societies in the real world? Setting out to scrutinise the external validity of previous findings, we envisage the white-supremacist rally in Charlottesville in 2017 as such a stark reminder of existent racism, which rendered American society conscious of its structural discrimination. More specifically, we test whether there are any substantial changes in implicit racial bias in a broad sample of US and UK online media by comparing the levels of pre- and post-Charlottesville bias. The sample we use contains 97,542 articles from 47 media outlets, combines tabloid and broadsheets, online bogs and satirical magazines, and spans from the extreme right to the left of the political spectrum. To operationalise racial bias we resort to the literature on the logic of Implicit Association Tests (IAT) that were developed to empirically test racial bias via word group associations. Drawing on this idea, we measure the association between specific word groups in written media. We fit word embedding models to pre- and post Charlottesville media samples and calculate cosine similarities between lists of words denoting black or white actors as well as positive and negative adjectives for character traits. This allows us to quantify the change in media bias towards African-Americans before and after the rally, compared to the bias towards white actors. We find that after the rally, there is no considerable change in positive bias towards white or black actors, while post-Charlottesville African-Americans are associated considerably less with negative character traits. Our findings suggest that media bias towards marginalized groups can temporarily shift after exogenous shocks such as the Charlottesville rally.

2 Interventions to Reduce Racial Bias

In spite of a general empirical tendency showing that racial bias is gradually dwindling since the 1960s (Gaertner and Dovidio 1986, Schuman et al. 1997), African-Americans are still suffering from structurally unequal treatment, such as poor quality interactions (McConnell and Leibold 2001), limited employment opportunities (Bertrand and Mullainathan 2004) or smaller chances of receiving life-saving medical treatment (Green et al. 2007). Racially prejudiced behaviour is believed to originate from implicit biases (Devine 1989, Gaertner and Dovidio 1986), which produces discriminatory behaviour (McConnell and Leibold 2001). These biases are reproduced in inter-personal interactions but also in collective means of communication, such as newspaper articles, fake news or blog entries. Past research has shown that media content produces negative dispositions towards such minority groups (Boomgarden and Vliegenthart 2009) and can be amplified by respective exogenous shocks (Czymara and Schmidt-Catran 2017). Results from earlier studies also indicate that racist bias is not static, but can be reduced temporarily or even in the long-term (Galinsky and Moskowitz 2000, Devine et al. 2012). For an individual to reduce their racial biases, the first step is to grow conscious of their bias, which is linked to the evocation of concern and guilt (Devine 1989), which motivates self-regulation to discontinue biased behaviour (Amodio, Devine, and Harmon-Jones 2007). Long-term de-biasing effects were achieved if this was coupled with bias education programmes designed to evoke general concern about implicit biases (Devine et al. 2012). Presenting an individual with feedback of their racial bias, thereby rendering them conscious of their bias and evoking concern about the racist biases held, can thus pave a way into decreasing racial prejudice.
Implicit association tests (IAT) is a method developed to lay bare socially significant associative structures and can be used to measuring evaluative associations that underlie implicit, e.g. racially biased, attitudes (Greenwald, McGhee, and Schwartz 1998). In these IAT, in essence, participants are made to answer to certain words with other words, e.g. names perceived to be typically *white* or *black* have to be replied to with words that fall under the category *pleasant* or *unpleasant*. If for instance an association between an example of each of the categories *white* and *pleasant* is stronger, this indicates an underlying positive bias towards the category white. The logic of IAT has been used in the application of word embeddings to text to track stereotypes on gender and minorities (Garg et al. 2018). We make use of this application of word embeddings to quantify negative and positive sentiment towards African-Americans and white Americans.

3 Media Bias Pre- and Post Charlottesville

Our example looks at the case of persistent racism by the US population towards African-Americans. In this example, we gauge public opinion by a broad range of media sources and compare the implicit bias towards the black and white population before and after a march of white-supremacists who expressed their overtly racist stances. Doing this, we use the Charlottesville rally as an event that serves like a nation-wide intervention of conscious-rendering. In this argument we draw on the mechanisms from the social psychology literature, which is focused on the effects of de-biasing on indi-

viduals. Analogously to social psychologists, which examine the effects on a group of participants, we examine a potential effect of collective conscious-rendering on public opinion, for which we use a large sample of media outlets and popular news blogs as proxy. Although the consumption of biased news can inform racially biased or racist believes, our focus in this research preliminary lies on the media as a proxy for public opinion and debate in society.

On 11 August 2017 the *Unite the Right* rally took place in Charlottesville, Virginia. The march consisted largely of white men, who self-identified as alt-right, neo-Confederates, neo-fascists, neo-Nazis, white nationalists and supremacists. The marchers chanted racist and anti-Semitic slogans, carried swastika and torches. Although the Charlottesville rally was previously announced, the level of racist slander, violence and the homicide committed by a rally member on the early morning of 12 August came as an as-good-as external shock, laying bare the perilous racism and its violent potential existing in US society. The intensity of overt racism and violence must have also evoked general concern among (at least a large part) of the public and caused a nation-wide debate. The Charlottesville rally serves as a treatment of collective feedback and evocation of general concern. If the bias-breaking mechanisms put forward by the psychological literature were to hold in the US case, we would expect a neutralisation of biases towards black actors. Indicators of such a neutralisation would entail that we would find neither a substantially more positive nor a substantially more negative bias towards black people in the post-Charlottesville media. We could also expect a new biasing effect, in which in a post-Charlottesville world, there would be more negative bias towards Caucasians.

4 Data and Method

Our media sample uses 97,542 online newspaper articles and blog entries from a variety of US and UK sources, spanning the period of 10 May to 11 November 2017. The sample includes mainstream and well-established newspapers such as CNN or The Guardian, fake news blogs and newspapers and satire sources as well as hyper-partisan political outlets such as Breitbart (Horne et al. 2018). The British media was included into the sample as the UK also has a large population, which faces structural negative bias, but also to increase article numbers. We removed sources with a very low publication output and processed articles by common methods of automated text analysis (Grimmer and Stewart 2012). In this process, we also removed short articles with fewer than 50 terms. This eventually resulted in a news sample of predominantly right-wing media, which will only allow us to infer shifts in bias in more conservative-oriented public opinion. Future research should use a more diversified sample across the political spectrum. The sources and number of articles for our final sample can be found in Table 1.

Figure 1 further displays the term frequency of the most frequent words of all articles that stem from the post-rally sample and include the term *Charlottesville*. While terms related to political actors and the rally itself appear very frequently, it is also apparent that the post-Charlottesville reports frequently discuss racism and violence. For this reason, we should unsurprisingly find an increased association between ethnicity of actors and terms such as *violence*, such that Caucasians are more closely associated with these terms after the rally.

Source	No. of articles
True Pundit	7313
Washington Examiner	6382
Breitbart	5964
BBC	5183
Drudge Report	4427
CNN	3391
New York Post	2920
The Huffington Post	2705
National Review	2610
Salon	2485
The Daily Beast	2321
RedState	2310
Politicus USA	2226
CBS News	2071
Daily Mail	2035
The Gateway Pundit	2017
Bipartisan Report	2011
CNBC	1870
TheBlaze	1757
Freedom Daily	1756
Vox	1711
The New York Times	1706
New York Daily News	1671
NPR	1589
RT	1585
The Political Insider	1523
NewsBusters	1498
ThinkProgress	1342
The Guardian	1332
USA Today	1294
Conservative Tribune	1286
Infowars	1282
Natural News	1259
The Duran	1211
The Atlantic	1208
The Daily Caller	1205
CNS News	1151
Counter Current News	1133
Fox News	1061
Media Matters for America	1050
The D.C. Clothesline	1036
PBS	1033
Talking Points Memo	1030
Yahoo News	1012
Daily Kos	997
The Right Scoop	925
The Conservative Tree House	658

Table 1. News sources and report counts.

However, in this work, we instead scrutinize whether the Charlottesville rally affected differences between ethnic groups not for terms directly related to racism and violence, but instead for positive and negative character traits. This allows us to examine racial bias for terms that are not directly related to the rally. We argue that we should only see an increased association between positive or negative character traits and ethnic actors post-Charlottesville if the rally affected the racial bias.

Thus, a stronger association of e.g. *black* and the word *friendly* after the rally would denote an increase of positive racial bias towards Africa-Americans. To operationalise both ethnic groups, we compile two sets of dictionaries, i.e. list of terms. We compile this list drawing on previous work that uses terms related to either African-Americans (black) or Caucasians (white) (Kozlowski, Taddy, and Evans 2018).

We then select terms that occur in our corpus using pre-existing dictionaries for character traits commonly perceived as positive (e.g. friendly) and negative (e.g. unreliable) (Gunkel 2019).

We examine bias towards African Americans with an automated text analysis approach relying on *doc2vec*, a recent variant of word embeddings (Mikolov et al. 2013, Le and Mikolov 2014). In comparison to bag of words approaches, which do not take into account the syntax of language (Grimmer and Stewart 2012), word embeddings can capture more complex semantic relations. Given enough training data this allows word embedding models to solve analogy tasks. For instance, the analogy problem *man is to woman as king is to ?* can be solved with the arithmetic operation *king - man + woman* applied to vectors learned from an embedding model, which would return the result *queen* (Kozlowski, Taddy, and Evans 2018). This powerful method is increasingly acknowledged by scholars and is, for instance, utilised to study the development of societal stereotypes (Garg et al. 2018) that are captured by algorithms trained on textual data (Bolukbasi

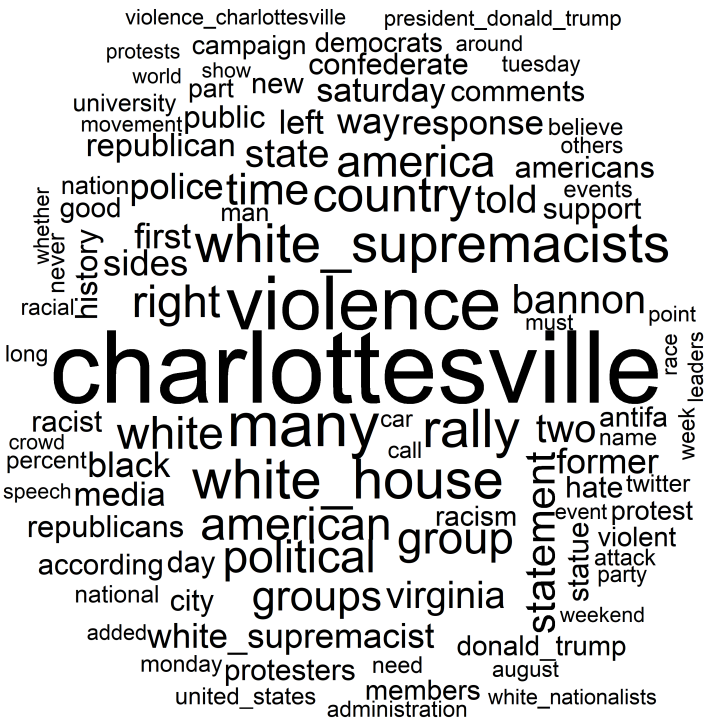

Figure 1. Word cloud of term frequency of all post-Charlottesville articles on the rally.

et al. 2016). In our paper, we instead focus on short- and mid-term developments of bias rather than stereotypes. We begin by training separate *doc2vec* models on articles published in two time periods, one three months before and three months after the rally. For each period, we train 20 models on bootstrapped samples of articles from the respective periods. The articles used to train each model are drawn at random with replacement. This allows us to not only examine biases before and after the Charlottesville rally but also to quantify uncertainty in our estimates (Kozlowski, Taddy, and Evans 2018). For each model, we project ethnicity on a polarity scale (Caucasian vs. African-American dimension) based on the ethnic dictionary. We then compute cosine similarities between ethnicity and positive as well as negative traits. This enables us to analyse the change in media bias towards African-Americans in comparison to Caucasians, as well as before and after the rally. In terms of research design, we are aware of the limitations of the causal claims we can make. We cannot randomly assign the *treatment* of the rally to a subset of the news outlets and therefore cannot control for possible confounders. However, our design still allows us to determine shifts in semantic associations between ethnicity and character traits before and after the rally.

5 Results

To examine whether the Charlottesville rally could have affected biases towards African-Americans, we visualise the results of the word embedding models using cosine similarities in Figure 2 (negative traits) and Figure 3 (positive traits). Due to space constraints we only visualise ten terms for each figure although the findings for negative and positive traits also apply to the remaining terms in our more comprehensive dictionaries.

Both Figures include the ethnicity dimension, where the left-hand side (negative values) is associated with Caucasian and the right-

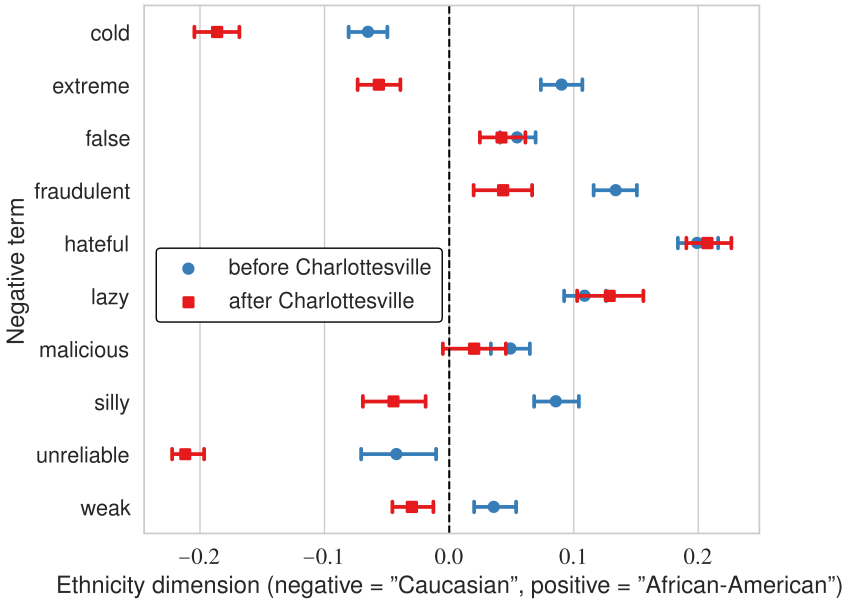

Figure 2. Cosine similarities with bootstrapped 90% intervals between ethnicity dimension and negative character traits.

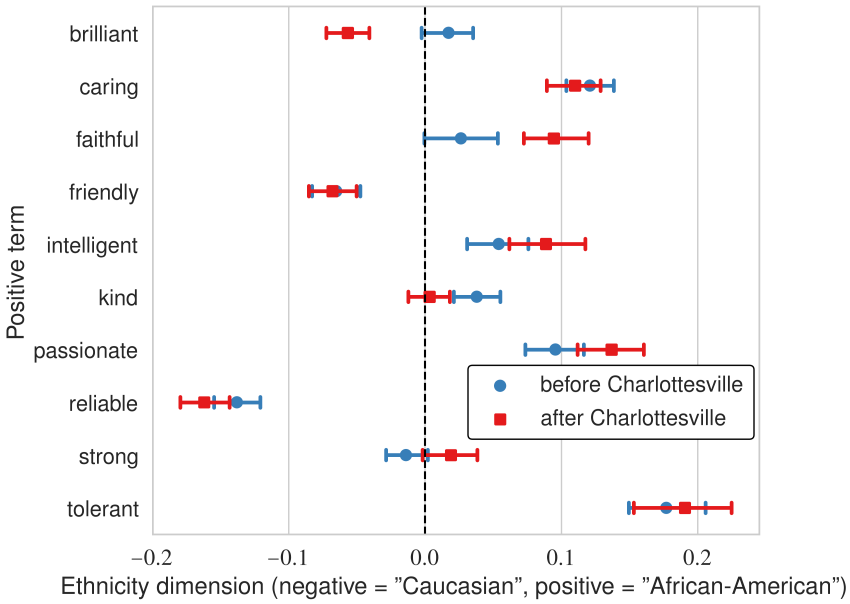

Figure 3. Cosine similarities with bootstrapped 90% intervals between ethnicity dimension and positive character traits.

hand side (positive values) is associated with African-American. Values for each term denote cosine similarities between the ethnicity dimension and the character trait. To give an example for the interpretation of the results, Figure 2 includes cosine similarities between the ethnicity dimension and negative term *silly*, both before and after the Charlottesville rally. Before Charlottesville *silly* was semantically closer to African-Americans. After the rally, the negative racial bias shifts, so that also the term *silly* is more closely related to Caucasians. This change in overall negative bias towards white actors in the post-rally

sample can also be observed with regards to the character traits in our dictionary that are not displayed in Figure 2. Unlike the negative character traits, the shift in positive bias between pre- and post-rally media reporting is minor, as can be seen in Figure 3. Bootstrapped intervals for the positive trait *intelligent* and other terms before and after the rally overlap, indicating minor or no differences.

Altogether, our findings do not suggest that there are any meaningful changes for associations between ethnicity and positive character traits. The similarities for positive terms in articles published after the Charlottesville rally are mostly in line with the similarities from articles published before the rally. However, the change in bias for negative terms is substantially larger. African-Americans are associated substantially less with negative character traits post-Charlottesville, while white actors are semantically closer to negative traits. These results suggest that at least for a short time period after the Charlottesville rally, articles in the digital media contained fewer associations between negative traits and African-Americans, i.e. a decrease in negative bias towards black people.

6 Conclusion

In this paper we set out to scrutinise whether the white-supremacist rally in Charlottesville in 2017 could have brought about any shift of positive and negative racial biases towards black and white Americans in the media. This research draws on social psychological concepts and mechanisms, such as IATs and bias-breaking interventions, and tests them on the aggregate level of American media. Theoretically, we could expect a de-biasing effect after the rally, meaning that there would be no notable difference between positive and negative associations of white and black people. Such a neutralisation could be the result of a successful bias-breaking intervention that renders individuals, or in our case the media, conscious of their previously held negative bias towards African-Americans and positive bias towards white Americans. To measure racial bias we compare the pre- and post-rally similarities of associations between ethnic terms and words for character traits. We find that there is no difference in pre- and post-rally media samples for positive associations, meaning Charlottesville did not seem to have had an effect of positive bias towards whites and blacks. However, we can observe a substantive overall shift in negative media bias towards black and white people after the rally. After the march, black people are associated less with negative terms than prior to the rally, while white actors are associated more with negative character traits post-rally. This holds despite of the fact that our sample comprises more right-wing than left-wing or centrist sources, so that we would expect to see similar but stronger effects in a more balanced sample of American news. Future research could build upon our work by looking into how and whether different segments of news outlets across the political spectrum adapt their implicit bias after such politically disrupting events. Scholars could also use the application of word embeddings in an experimental or quasi-experimental framework to isolate clear causal effects and test the theory of bias-breaking on the aggregate level of societies. Despite the shortcomings of our work, we seek to contribute to the literature on marginalized groups in digital societies, showing that media biases towards marginalized groups can temporarily decrease in the light of exogenous shocks.

7 Acknowledgments

Earlier versions of this paper were presented at the GESIS Summer School on Methods for Computational Social Science and the European Symposium on Societal Challenges in Computational Social Science. We thank the participants at these events for their useful comments. In particular, we thank Damian Trilling who initiated this project and James Evans for his many helpful suggestions.

8 References

Amodio, David M., Patricia G. Devine, and Eddie Harmon-Jones (2007). "A dynamic model of guilt: Implications for motivation and self-regulation in the context of prejudice". In: Psychological Science 18, E542–E530.

Bertrand, Marianne and Sendhil Mullainathan (2004). "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination". In: American Economic Review 94.4, pp. 991–1013.

Bolukbasi, Tolga et al. (2016). "Man is to Computer Programmer As Woman is to Homemaker? Debiasing Word Embeddings". In: Proceedings of the 30th International Conference on Neural Information Processing Systems. NIPS'16. USA: Curran Associates Inc., pp. 4356–4364.

Boomgarden, Hajo G. and Rens Vliegenthart (2009). "How news content influences anti-immigration attitudes: Germany, 1993-2005". In: European Journal of Political Research 48.4, pp. 516–542.

Czymara, Christian S and Alexander W Schmidt-Catran (2017). "Refugees Unwelcome? Changes in the Public Acceptance of Immigrants and Refugees in Germany in the Course of Europe's 'Immigration Crisis'". In: European Sociological Review 33.6, pp. 735–751.

Devine, Patricia G. (1989). "Stereotypes and prejudice: Their automatic and controlled components." In: Journal of Personality and Social Psychology 56.1, pp. 5–18.

Devine, Patricia G. et al. (2012). "Long-term reduction in implicit race bias: A prejudice habit-breaking intervention". In: Journal of Experimental Social Psychology 48.6, pp. 1267–1278.

Gaertner, Samuel and John F. Dovidio (1986). "The aversive form of racism". In: Prejudice, discrimination, and racism. Orlando: Academic Press, pp. 61–89.

Galinsky, Adam D. and Gordon B. Moskowitz (2000). "Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism." In: Journal of Personality and Social Psychology 78.4, pp. 708–724.

Garg, Nikhil et al. (2018). "Word embeddings quantify 100 years of gender and ethnic stereotypes". In: Proceedings of the National Academy of Sciences 115.16, E3635–E3644.

Green, Alexander R. et al. (2007). "Implicit Bias among Physicians and its Prediction of Thrombolysis Decisions for Black and White Patients". In: Journal of General Internal Medicine 22.9, pp. 1231–1238.

Greenwald, Anthony G, Debbie E McGhee, and Jordan L K Schwartz (1998). "Measuring individual differences in implicit cognition: the implicit association test." In: Journal of personality and social psychology 74.6, p. 1464.

Grimmer, Justin and Brandon M Stewart (2012). "Text as Data: The Promise and Pitfals of Automatic Content Analysis Methods for Political Texts". In: Political Analysis 21.617, pp. 267–297.

Gunkel, Patrick (2019). 638 Primary Personality Traits. URL: http://ideonomy.

mit . edu / essays / traits . html (visited on 02/08/2019).

Horne, Benjamin D. et al. (2018). "Sampling the News Producers: A Large News and Feature Data Set for the Study of the Complex Media Landscape". In: arXiv preprint 1803.10124.

Kozlowski, Austin C., Matt Taddy, and James A. Evans (2018). "The Geometry of Culture: Analyzing Meaning through Word Embeddings". In: arXiv preprint 1803.09288.

Le, Quoc V. and Tomas Mikolov (2014). "Distributed Representations of Sentences and Documents". In: arXiv preprint 1405.4053.

McConnell, Allen R. and Jill M. Leibold (2001). "Relations among the Implicit Association Test, Discriminatory Behavior, and Explicit Measures of Racial Attitudes". In: Journal of Experimental Social Psychology 37.5, pp. 435–442.

Mikolov, Tomas et al. (2013). "Distributed Representations of Words and Phrases and their Compositionality". In: Advances in Neural Information Processing Systems 26. Curran Associates, Inc., pp. 3111–3119.

Schuman, Howard et al. (1997). Racial attitudes in America: Trends and interpretations. Cambridge, MA: Harvard University Press.