Zusammenhang zwischen sportlichem Erfolg und Darstellung bei Wikipedia im Vergleich zwischen Sportlern und Sportlerinnen

Josefine Busch, Lisa Hillebrand, Flip Jansen, Daniela Tumbrägel

Grundlagen sozialer Netze
Prof. Dr. Gefei Zhang
Wintersemester 2018/2019

Hochschule für Technik und Wirtschaft
Berlin

Zusammenfassung

—

1 Problemformulierung

Die Motivation für unser Projekt ergab sich aus dem Fall von Donna Strickland, die im Jahr 2018 - zusammen mit Gérard Mourou und Arthur Ashkin - einen Nobelpreis im Bereich der Laserphysik erhalten hat [11]. Sie ist die dritte Frau, der je ein Physik-Nobelpreis verliehen wurde und erst nach der Anerkennung bekam sie einen Wikipedia-Artikel (cf. Washington Post). Ihre männlichen Kollegen sind in der Online-Enzyklopädie schon seit einigen Jahren vertreten. Einige Monate zuvor wurde ein Eintrag über Strickland abgelehnt, da sie von einem der Wikipedia-Editoren als nicht wichtig genug erachtet wurde [20]. Neben zahlreichen anderen Beispielen, zeigt dies die geschlechtsspezifische Diskriminierung auf Wikipedia und insbesondere die Ausgrenzung von Frauen in der Wissenschaft auf (cf. Guardian).

Geplant war, durch unser Projekt, numerisch und graphisch darzustellen, dass Frauen in der Wissenschaft auf Wikipedia schlechter repräsentiert werden als Männer. Bei der Suche nach effizienten Datenquellen- und -sätzen ergab sich allerdings die Problematik, dass Frauen in der Wissenschaft insgesamt so wenig vertreten sind, dass man keinen sinnvollen Vergleich zwischen Frauen und Männern herstellen kann. [TBE: Die Daten sind einfach nicht da, was ein ganz anderes gesellschaftskritisches Fass aufmacht].

Wir entschieden uns deshalb für einen anderen Themenbereich, in dem Frauen und Männer gleichermaßen vertreten sind: Sport. Um die Sache entsprechend einzugrenzen, wählten wir die Olympischen Sommerspiele 2016 in Rio de Janeiro.

Im Hinblick auf die größtenteils gleichmäßige Anzahl von beiden Geschlechtern, untersuchen wir das Verhältnis von sportlichem Erfolg und der entsprechenden Darstellung auf Wikipedia. Motiviert durch die oben genannten Ereignisse, bildet sich unsere These, dass Frauen bei gleichzusetzender Leistung im Vergleich zu Männern auf Wikipedia unzureichend repräsentiert werden. Im Detail differenzieren wir dabei die mittlere Länge und die mittlere Anzahl der redaktionellen Änderungen der Wikipedia-Artikel der Sportler*innen der Olympischen Sommerspiele im Jahr 2016.

Theorie

Wikipedia

Wikipedia wurde im Jahr 2001 gegründet [18]. Mittlerweile gibt es sprachenübergreifend über 46 Millionen Artikel [21]. Von rund 35 Millionen registrierten Nutzer:innen sind jedoch nur etwa 124.000 monatlich aktiv [23]. Wikipedia ist zum Zeitpunkt des Verfassens dieses Texts auf dem fünften Platz der weltweit populärsten Websites [1] und hat damit eine enorme Reichweite. Jede Person kann zu Wikipedia beitragen und auch ohne Registrierung Artikel bearbeiten [22]. Das Erstellen neuer Artikel ist nur nach vorheriger Registrierung möglich. Jeder Artikel verfügt über eine Versionsgeschichte, in der vorherige Artikelversionen sowie die Anmerkungen der Autor:innen und Administrator:innen gespeichert werden.

Gender Bias

Ein Bias ist eine Verzerrung bzw. ein systematischer Fehler [24]. Als Gender Bias bezeichnet man eine Bevorzugung oder Parteilichkeit gegenüber einer oder mehreren Personen ausschließlich wegen des Geschlechts. Es sollen nun einige Erkenntnisse hinsichtlich des Gender Bias im Sport zusammengefasst und und daraus die Hypothesen abgeleitet werden.

1.0.1 Gender Bias und Wikipedia

In mehreren Studien konnte gezeigt werden, dass ein Gender Bias sowohl in der Zusammensetzung der Autorenschaft bei Wikipedia als auch hinsichtlich der inhaltlichen Gestaltung der Artikel besteht. Unter den Autoren sind bestimmte Personengruppen deutlich unterrepräsentiert. In Nutzerbefragungen wurde erfasst, dass zwischen 84% und 90% aller Wikipidia-Autor:innen männlich sind [17, 4]. Hierzu muss jedoch angemerkt werden, dass die von der Wikimedia Foundation selbst durchgeführten Erhebungen Opt-in-Befragungen sind, die zu Verzerrungen führen können. So ist vorstellbar, dass weniger Frauen als Männer an der Befragung teilnehmen, sie aber eigentlich unter den Autoren deutlich häufiger vertreten sind. Hill und Shaw [7] überprüfen genau diese Kritikpunkte und berechnen einen korrigierten Wert für den Anteil Autorinnen, der jedoch auch nach Korrektur bei lediglich 16.1% liegt. Es lässt sich also zusammenfassen, dass nach jetzigem Stand der Kenntnis deutlich weniger Frauen als Männer zu Wikipedia beitragen.

Auf inhaltlicher Ebene untersuchten verschiedene Studien das Vorliegen eines Gender Bias. Wagner u. a. [16] analysierten Wikipedia-Artikel über bekannte Personen in vier Parametern: Abdeckungsbias (coverage bias), Strukturbias (structural bias), Lexikalischer Bias (lexical bias) und Sichtbarkeitsbias (visibility bias). Der Abdeckungsbias beschreibt, über wie viele Personen Wikipedia-Artikel bestehen. So könnte eine Annahme sein, dass über bekannte Frauen weniger Artikel verfasst werden als über bekannte Männer. Mit dem Strukturellen Bias werden Unterschiede beispielsweise in der Verlinkungsstruktur beschrieben. Ein Bias würde bestehen, wenn Frauenartikel mehr Verlinkungen auf Artikel über Männer enthalten als umgekehrt. Ein Lexikalischer Bias umfasst Unterschiede im Vokabular, mit dem Männer und Frauen beschrieben werden. Ein Beispiel für einen Bias wäre, dass in Artikeln über Frauen mehr Vokabular mit Beziehungs- oder Familienbezug verwendet wird. Der Sichtbarkeitsbias gibt an, ob es Unterschiede zwischen den Geschlechtern in der Häufigkeit gibt, mit der Artikel auf der Startseite von Wikipedia platziert werden. In ihrer Studie konnten Wagner u. a. [16] Evidenz für den Strukturellen und den Lexikalischen Bias finden. Artikel über Frauen verlinken häufiger auf Artikel von Männern als umgekehrt. Auf lexikalischer Ebene konnten die Autoren zeigen, dass in Artikeln über Frauen mehr Vokabular mit Bezug zu Familie und romantischen Beziehungen verwendet wird als in Artikeln über Männern. Ein Abdeckungsbias konnte nicht gezeigt werden; bekannte Frauen und Männer, die für die Studie ausgewählt wurden, waren quantitativ gleich gut vertreten.

In einer Studie von Graells-Garrido, Lalmas und Menczer [4] fand sich hingegen, dass der Gender Bias auf Wikipedia mit Struktur und Inhalt von Artikeln zusammenhängt. In der Artikellänge besteht ein signifikanter Unterschied, wobei Artikel über Frauen im Mittel kürzer sind als Artikel über Männer. Die Effektstärke fiel lediglich gering aus.

Die vorliegende Untersuchung fokussiert auf den Abdeckungsbias und untersucht die Artikellänge sowie die Anzahl der Editierungen.

1.0.2 Gender Bias im Sport

In einigen Sportarten erhalten Sportler mehr Aufmerksamkeit als Sportlerinnen, was sich in den Umsätzen und schlussendlich auch in den Gehältern niederschlägt. So liegt das Durchschnittsgehalt in der Fußball-Bundesliga der Männer bei rund 1,5 Millionen Euro [6], während das Durchschnittsgehalt der Frauen bei 43.000 Euro pro Jahr liegt [13]. Zahlreiche Studien haben sich mit der Medienberichterstattung über Sportler und Sportlerinnen befasst. Zum einen gibt es Sportarten, in denen weitaus mehr Zeit der TV-Berichterstattung für die männlichen Sportler aufgewendet wird - Fußball oder Basketball sind hier nur zwei Beispiele. In der NBC-Berichterstattung über die Olympischen Spiele 2008 wurde 54,2% der Übertragungszeit über Sportler berichtet und 44,8% der Zeit über Sportlerinnen [2]. Trolan [14] fassten Erkenntnisse anderer Studien zusammen und kommen zu dem Schluss, dass Frauen in der Medienberichterstattung sowohl quantitativ unterrepräsentiert sind als auch qualitativ anders berichtet wird. Die Leistungen von Frauen werden trivialisiert, indem ein stärkerer Fokus auf das Äußere von Sportlerinnen gelegt wird als auf ihre Leistungen [5, 15]. Jones, Murrell und Jackson [8] untersuchten die Berichterstattung über die Olympischen Spiele 1998 und fanden, dass über Sportlerinnen mehr berichtet wird, wenn sie einer Sportart nachgehen, die als weniger maskulin beurteilt wird. Billings und Angelini [2] untersuchten, wie Sportkommentator*innen Erfolg und Misserfolg bei Sportler*innen attribuieren. Im Falle von Erfolg führten sie diesen bei Sportlern signifikant häufiger auf deren überlegene körperliche Stärke zurück. Zusammenfassend lässt sich also sagen, dass es quantitative und qualitative Unterschiede in der Berichterstattung über Sportler und Sportlerinnen gibt. Daraus abgeleitet ergibt sich die Annahme, dass sich dies auch in der Aufmerksamkeit niederschlägt, die Sportler*innen bei Wikipedia erfahren.

Hypothesen

Es ergeben sich folgende Hypothesen:

Hypothese 1. Wikipedia-Artikel über Sportler sind signifikant länger als Wikipedia-Artikel über Sportlerinnen.

Hypothese 2. Wikipedia-Artikel über Sportler wurden signifikant häufiger editiert als Wikipedia-Artikel über Sportlerinnen.

Diese Untersuchung überprüft die Hypothesen anhand von acht ausgewählten Wettkämpfen bei den Olympischen Spielen in Rio 2016.

Methode //Josefine, Busch 560106

Erhebungsart

Zur Beschaffung der Daten stehen uns zwei Möglichkeiten zur Verfügung. Auf der offiziellen Webseite der Olympischen Kommission sind die Namen der Teilnehmenden der vergangenen Spiele veröffentlicht.[12] Der Vorteil dieser Daten ist, dass sie bereits in weibliche und männliche Teilnehmer unterteilt sind. Alternativ sind auch direkt auf Wikipedia Listen der Teilnehmenden zu finden. [19] Die Wikipedia Listen beinhalten direkte Verlinkungen zu den jeweiligen Seiten der Sportler*innen, allerdings ist hier nicht mehr ohne Weiteres auszulesen, welchem Gender die Person angehört. Um diese Information zu bekommen müsste der Fließtext auf die verwendeten Pronomen untersucht werden. Auch andere Informationen sind bei Wikipedia nicht mehr so leicht auslesbar wie aus der Liste von Olympic.org. Besonders sind die Sportler*innen nach

dem Alphabet geordnet nicht nach dem erreichten Rang, dieser müsste also ebenfalls aus dem Fließtext ausgelesen werden.

Die Liste der Olympischen Spiele 2016 von Olympic.org enthält strukturierte Daten in Form der folgenden Spalten:

- Year
- Sport
- Discipline
- Event
- EventGender
- Phase
- PhaseXX
- Unit
- UnitXX
- Names
- Gender
- NOC
- Teammembers
- Rank
- Results
- Medal
- Adversary
- Adversary NOC

Die für uns interessanten Daten hieraus sind die Spalten Sport, Discipline, Event, Names und Gender. Die Länge der Artikel, gemessen an der Anzahl der Wörter, und die Anzahl der Editierungen der einzelnen Artikel müssen erst berechnet beziehungsweise von Wikipedia direkt abgefragt werden. Diese Daten liegen in unstrukturierter Form vor.

Stichproben

Da nicht in allen Sportarten Männer und Frauen in vergleichbaren Disziplinen antreten und in manche Disziplinen Teams gegeneinander antreten, wählen wir acht Disziplinen aus unterschiedlichen Sportarten aus, welche unsere Anforderungen erfüllen. Die Anforderungen sind, dass die Disziplinen für Männer und Frauen gleich sind bzw. gleich bezeichnet sind und dass sie keine Teamdisziplinen sind. Die ausgewählten Sportarten sind:

- Turmspringen (10m)

- Bogenschießen
- Fechten (épée)
- Moderner Fünfkampf
- Stabhochsprung
- Schwimmen (100m Freestyle)
- Radfahren
- Athletik (100m)

Um die Artikel von Wikipedia zu bekommen und von diesen die Länge in Wörtern verwenden wir ein Modul von GitHub Benutzer Jonathan Goldsmith [3]. Mit diesem Modul und den Titeln der Wikipedia Seiten können die Artikel abgerufen werden. Danach musst nur noch die Wörteranzahl ermittelt werden. Mit Stichproben soll geprüft werden, dass die korrekten Artikel zurück gegeben und gezählt werden.

Für die Zwischenspeicherung der Daten im Jupyter Notebook verwenden wir Pandas dataFrames. Außerhalb des Programms sind die Daten in CSV-Dateien abgelegt. Für die grafische Darstellung im Notebook verwenden wir außerdem Matplotlib[9] und Seaborn[10].

Auswertungsmethoden

Mit Histogrammen zu den unterschiedlichen Disziplinen wollen wir die Verteilung der Anzahl der Wörter für männliche und weibliche Sportler vergleichen und gegebenenfalls einen T-Test durchführen. Zur weiteren Untersuchung der Extremwerte, des Mittelwerts und des Median sollen Boxplots und Balkendiagramme zu den verschiedenen Disziplinen erstellt werden.

Data Privacy

——————Aus Prof. Zhangs Skript ——————

Methodologie / Untersuchung Datenbeschaffung Datenpräparation Exploration/Modell Planung (Ziel: Hypothese) Model-Erstellung (Ziel: Test Hypothese, ggfs. Anpassung des Models)

Wie gehen Sie vor und warum? Welche Werkzeuge nutzen Sie? § Datenbeschaffung (Qualität und Quantität der Daten mit Darstellung weiterer Quellen / Erhebungsmethoden?) » Erhebungsart » Stichprobe » Auswertungsmethoden § Ethische Kriterien berücksichtigen (vgl. Data Privacy) » Missbrauch personenbezogener Daten (Anonymisierung, Pseudonymisierung)

2 Durchführung // Jansen, Flip 558059

Ausgehend von der Wikipedia-Überblicksseite [19]haben wir rekursiv alle Sportarten sowie auf der untersten Ebene die Links der einzelnen Sportler*innen einer Sportart auf Wikipedia gesammelt und in eine CSV-Datei geschrieben.

Um mögliche Fehlern beim Sammeln der Daten von der Olympic.org - Seite zu vermeiden, haben wir das Olympic Studies Centre angeschrieben, die uns eine Excel-Tabelle mit den ausführlichen Ergebnissen der Olympischen Sommerspiele 2016 zur Verfügung gestellt haben.

Um die Wikipedia-Links mit den von uns benötigten Daten der Ergebnistabelle zu verbinden, mussten wir die unterschiedlichen Namensformate anpassen und Sonderzeichen umwandeln. So setzt zum Beispiel Wikipedia bei Namensgleichheit von in Wikipedia erfassten Personen eine Kategoriebezeichnung ans Ende des Namens: "(diver)".

Trotzdem konnten wir einige Links nicht automatisiert zuordnen, da es zum einen in den Ergebnistabelle Rechtschreibfehler in den Namen gibt, zum anderen, weil gerade bei Namen aus Sprachgebieten mit nicht-lateinischer Schrift unterschiedliche !!!Konvertierungen!!!

3 Ergebnisse

4 Diskussion

Literatur

[1] Alexa Internet Inc. *Alexa Top 500 Global Sites*. URL: https://www.alexa.com/topsites (besucht am 09. 01. 2019).

[2] Andrew C. Billings und James R. Angelini. „Journal of Broadcasting & Gendered Profiles of Olympic History : Sportscaster Dialogue in the 2008 Beijing Olympics". In: *Journal of Broadcasting & Electronic Media* 54.October 2011 (2008), S. 37–41. URL: http://www.tandfonline.com/doi/abs/10.1080/08838150903550352.

[3] GitHub.com. *GitHub - goldsmith/Wikipedia: A Pythonic wrapper for the Wikipedia API*. URL: https://github.com/goldsmith/Wikipedia (besucht am 09. 01. 2019).

[4] Eduardo Graells-Garrido, Mounia Lalmas und Filippo Menczer. „First Women, Second Sex: Gender Bias in Wikipedia". In: (2015). URL: http://arxiv.org/abs/1502.02341%7B%5C%%7D0Ahttp://dx.doi.org/10.1145/2700171.2791036.

[5] John Harris. „The Image Problem in Women's Football". In: *Journal of Sport and Social Issues* 29.2 (2005), S. 184–197. URL: http://journals.sagepub.com/doi/10.1177/0193723504273120.

[6] Nick Harris. „Global Sports Salaries Survey 2015". In: (2015). URL: http://www.globalsportssalaries.com/GSSS%202015.pdf.

[7] Benjamin Mako Hill und Aaron Shaw. „The Wikipedia Gender Gap Revisited: Characterizing Survey Response Bias with Propensity Score Estimation". In: *PLoS ONE* 8.6 (2013), S. 1–5.

[8] Ray Jones, Audrey J. Murrell und Jennifer Jackson. *Pretty Versus Powerful in the Sports Pages*. 1999. URL: http://journals.sagepub.com/doi/10.1177/0193723599232005.

[9] Matplotlib. *Matplotlib: Python plotting*. URL: https://matplotlib.org/ (besucht am 09. 01. 2019).

[10] Michael Waskom. *seaborn - statistical data visualization*. URL: https://seaborn.pydata.org/ (besucht am 09. 01. 2019).

[11] Nobel Media AB. *The Nobel Prize in Physics 2018*. URL: https://www.nobelprize.org/prizes/physics/2018/summary/ (besucht am 09. 01. 2019).

[12] Olympic.org. *Olympic Results, Gold Medalists and Official Records*. URL: https://www.olympic.org/olympic-results (besucht am 09. 01. 2019).

[13] Statista. *Spielergehälter in Frauen-Fußballligen Saison 2017/2018*. URL: https://de.statista.com/statistik/daten/studie/819734/umfrage/spielergehaelter-in-frauen-fussballligen/ (besucht am 10. 01. 2019).

[14] Eoin J. Trolan. „The Impact of the Media on Gender Inequality within Sport“. In: *Procedia - Social and Behavioral Sciences* 91 (2013), S. 215–227. URL: http://linkinghub.elsevier.com/retrieve/pii/S1877042813025512.

[15] John Vincent. „Game, Sex, and Match: The Construction of Gender in British Newspaper Coverage of the 2000 Wimbledon Championships“. In: *Sociology of Sport Journal* 21.4 (2004), S. 435–456. URL: http://journals.humankinetics.com/doi/10.1123/ssj.21.4.435.

[16] Claudia Wagner u. a. *It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia*. 2015. URL: http://arxiv.org/abs/1501.06307.

[17] Wikimedia Foundation. *Community Engagement Insights 2018 Report*. URL: https://meta.wikimedia.org/wiki/Community%7B%5C_%7DEngagement%7B%5C_%7DInsights/2018%7B%5C_%7DReport (besucht am 09. 01. 2019).

[18] Wikimedia Foundation. *Wikipedia Timeline*. URL: https://meta.wikimedia.org/wiki/Wikipedia%7B%5C_%7Dtimeline (besucht am 09. 01. 2019).

[19] Wikipedia. *Competitors at the 2016 Summer Olympics*. URL: https://en.wikipedia.org/wiki/Category:Competitors%7B%5C_%7Dat%7B%5C_%7Dthe%7B%5C_%7D2016%7B%5C_%7DSummer%7B%5C_%7DOlympics (besucht am 09. 01. 2019).

[20] Wikipedia. *Donna Strickland - Wikipedia (old revision)*. URL: https://en.wikipedia.org/w/index.php?title=Draft:Donna%7B%5C_%7DStrickland%7B%5C&%7Doldid=842614385 (besucht am 09. 01. 2019).

[21] Wikipedia. *Size of Wikipedia*. URL: https://en.wikipedia.org/wiki/Wikipedia:Size%7B%5C_%7Dof%7B%5C_%7DWikipedia.

[22] Wikipedia. *Wikipedia Tutorial & Registration*. URL: https://en.wikipedia.org/wiki/Wikipedia:Tutorial/Registration (besucht am 09. 01. 2019).

[23] Wikipedia. *Wikipedians*. URL: https://en.wikipedia.org/wiki/Wikipedia:Wikipedians.

[24] Markus Antonius Wirtz, Janina Strohmer und Friedrich Dorsch. *Dorsch - Lexikon der Psychologie*. 18., über. Wirtz, Markus Antonius. URL: https://www.hogrefe.ch/shop/dorsch-lexikon-der-psychologie-75944.html.