Gender Markers in Wikipedia Usernames

Björn Ross
University of Duisburg-Essen
Duisburg, Germany
bjoern.ross@uni-due.de

Marielle Dado
University of Duisburg-Essen
Duisburg, Germany
marielle.dado@uni-due.de

Maritta Heisel
University of Duisburg-Essen
Duisburg, Germany
maritta.heisel@uni-due.de

Benjamin Cabrera
University of Duisburg-Essen
Duisburg, Germany
benjamin.cabrera@uni-due.de

ABSTRACT

Wikipedia is used as a source of information by many people online. Research has, however, indicated that there is a gender gap in participation. This gap is problematic because it influences which topics are covered and how they are portrayed on Wikipedia. As an example of the gap, posts by women on talk pages are slightly less likely to receive a reply than posts by men. It is still unclear whether this difference is the result of women being treated differently on talk pages than men. One of the only cues available to Wikipedia users for guessing the author of a talk page post's gender is their username, i.e. their pseudonym on the platform. We therefore examined whether users with obviously female names receive fewer replies than users with obviously male names. To this end, we developed and evaluated an automated method to determine whether a username contains obvious gender markers, based on machine learning. We then applied this algorithm to the entire data set of all Wikipedia talk page discussions. Contrary to our expectations, we find that users with clearly female names are slightly more likely to receive a reply than users with clearly male names. We also find that the fraction of users with a female name is much lower than the fraction of female users, suggesting that, unlike men, women using Wikipedia do not include contain obvious gender markers in their usernames. We discuss the implications of this result for Wikipedia and the Wikipedia research community.

KEYWORDS

Wikipedia, talk page, gender imbalance, user names

1 INTRODUCTION

The gender gap on Wikipedia, one of the most popular websites on the Internet, is widely documented. Research has demonstrated that not only are biographies of women more likely to be missing from Wikipedia [8], but they are also more likely to be characterised using personal information such as childbirth, marriage and family [1]. Women are also less likely than men to edit Wikipedia articles, especially those about topics that are traditionally male-dominated, such as science and engineering [4]. Recently, we investigated the gender gap on Wikipedia by determining how the likelihood of receiving a reply on a talk page depends on the author's gender [2]. We used three data sets (based on an XML dump of the English Wikipedia as of December 2017): all comments made by Wikipedia editors in talk pages, which were extracted using the application GraWiTas[3], the gender of these editors if they chose to disclose it (based on the gender indicated in the user preferences as well as user boxes in personal user pages), and articles belonging to several categories.

As you can see in Table 1, we found that female authors received fewer replies than male authors: specifically that they were 2.4% less likely to receive at least one reply (0.3418/0.3503 = 0.976). This gap is statistically significant, and it is more pronounced in traditionally male-dominated topics (e.g., engineering).

In this work-in-progress paper, we attempt to uncover the factors that might have led to this disparity by looking at the likelihood of receiving a reply based on gender markers in the editors' usernames (i.e., usernames with names commonly associated with men or women). Given the absence of physical presence on the Internet, usernames have become a way for Internet users to establish an online identity, often reflecting aspects of one's "real world" identity [10]. On Wikipedia talk pages, usernames are often the only cue for users to infer their fellow editors' genders. For instance, a recent study on a wiki-like platform found that women contributed less when there appeared to be no other visible female editors present, due to the lack of female-sounding usernames and the abundance of anonymous users, which were often assumed to be male editors [9]. Women have been found to be more likely to mask their gender identity on Internet forums by selecting numerical usernames, while men were comfortable with using their own names [6]. Out of the 4144 users in a study about StackOverflow Q&A, an online community for discussing technical themes, only 291 users had feminine identities based primarily on feminine markers in their usernames, compared to 2296 usernames with masculine identities and 1557 with no obvious gender markers [11]. This result is coupled with the finding that although women formulated more questions and provided a similar number of answers, men were engaged in forum discussions for longer periods of time. The researchers concluded that women tend to disengage faster and may be more likely to hide their gender identities to remain anonymous.

There appears to be a paradoxical situation on Wikipedia, whereby to encourage more women to contribute to Wikipedia, women should be able to see other women contributing to Wikipedia as well [9]. This paradox is exacerbated by the tendency of women to mask their gender identity in their usernames, presumably to avoid discrimination or bias. One study observed this bias in email correspondences between students and faculty members: faculty

Wiki Workshop, April 2018, Lyon, France
2018.

Table 1: Reply probability by gender

Top-level posts …	Number of posts
by male authors	1,641,282
— which received at least one reply	574,980 (35.03%)
by female authors	143,815
— which received at least one reply	49,163 (34.18%)

members replied more often to prospective students with male names than prospective students with female names [5]. This phenomenon is also present in the offline world: in a double-blind randomised experiment, professors in a science faculty rated applicants with a male-sounding name as more competent and 'hireable' than an identical applicant with a female name [7].

To explore how "feminine" or "masculine" Wikipedia editor usernames may be related to the reply rate to talk page posts, we created a classifier that determined whether a username belongs to a male or female editor based on certain gender markers. We hypothesise that talk page posts created by usernames with feminine markers have a lower reply rate than posts by usernames with masculine markers.

2 METHOD

To examine whether female gender markers in a username decrease the likelihood of their posts being replied to, we required a method that would automatically detect gender markers in internet usernames. The difference in reply rates between men and women (2.4%), although in the order of tens of thousands of comments on a website with several millions of comments such as Wikipedia, is rather small. If the difference in reply rates between people with obviously male names and people with obviously female names is of a similar size, it will be difficult to detect in a small sample. An automated tool, however, could easily classify the entire data set of Wikipedia users active on talk pages.

Machine learning was thus used to automatically infer how feminine or masculine a given username appears. More specifically, we used logistic regression to generate out-of-sample predictions of a user's gender based on their name. For training the classifier, we used the largest publicly available (and ethically acceptable) data set containing both usernames and gender information that we could find: a data set of around two million Last.fm profiles published on the Open Data platform Socrata[1]. Last.fm is a music website that allows people to track their listening history and recommends similar artists. At the time the data set was compiled (December 2012), the gender of a user, if they had chosen to disclose it, was publicly displayed on their profile page. In this data set, there are 1,460,833 users whose gender is known. Much like on Wikipedia, some Last.fm users use their real names or variations of it, while others choose a different name, including made-up ones.

[1]https://opendata.socrata.com/Business/Two-Million-LastFM-User-Profiles/5vvd-truf

Table 2: Evaluation of classification performance and accuracy of guesses

	Precision	Recall	F_1 score	Support
Logistic regression model				
Female	0.44	0.44	0.44	3011
Male	0.93	0.93	0.93	24681
Avg. / Total	0.88	0.88	0.88	27692
Humans (random sample)				
Female	0.19	0.57	0.29	72
Male	0.95	0.76	0.85	727
Avg. / Total	0.88	0.75	0.80	27692
Humans (stratified sample)				
Female	0.71	0.56	0.63	400
Male	0.64	0.77	0.70	400
Avg. / Total	0.68	0.67	0.66	800

Table 3: Top features indicating username gender

Male		Female	
n-gram	Coefficient	n-gram	Coefficient
'bruno'	33.85	'girl'	-53.64
'brendan'	30.53	'woman'	-41.00
'tephen'	29.47	'lady'	-39.89
'mike'	27.04	'emily'	-39.80
'joao'	26.73	'miss'	-33.80
'caio'	26.16	'ouise'	-33.40
'nathan'	25.64	'laura'	-32.89
'jesse'	25.29	'luiza'	-32.89
'jake'	25.08	'bruna'	-30.62
'fabio'	24.76	'grrl'	-30.47

For testing the classifier, we used the Wikipedia data set that we had collected (49,387 male and 5,996 female usernames)[2]. Character n-grams of arbitrary length were used as features (with tf–idf weighting).

Logistic regression was used in conjunction with L_1 regularisation because it produces probabilities and sparse parameter vectors which can be inspected manually. The regularisation parameter (C = 2) was chosen as a result of a grid search that was performed on 50% of the test data to optimise macro F_1. Table 2 summarises the classification performance as calculated using the remaining 50% of data.

The main goal of training the classifier, however, was not to maximise accuracy. Since the classifier should mimic the judgments on gender markers made by humans, the ideal classifier, for our purposes, should agree with the judgments that a human would make, and make the same 'mistakes' regarding the actual gender of the users.

[2]We also experimented with using the Wikipedia data for both training and testing, using cross-validation. The results indicated that the classifier was overfitting. Consequently, we collected the much larger Last.fm data set.

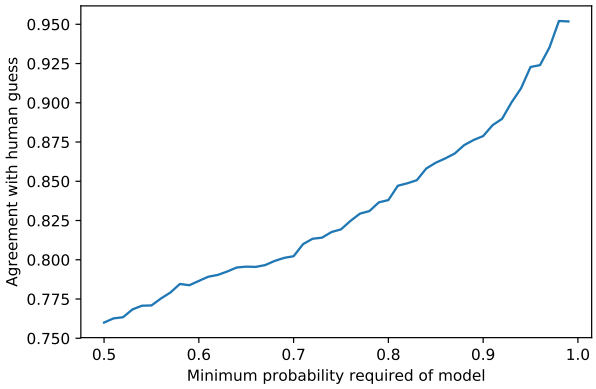

Figure 1: Relationship between the minimum probability required of the model to make a prediction and the fraction of humans that agree with the model's prediction

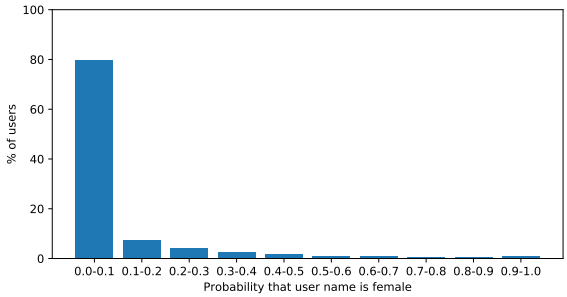

Figure 2: Frequency of male, unisex and female usernames on Wikipedia

To examine how the final model arrives at its decisions, the top ten features indicating male or female gender are reported in Table 3. The top n-grams are either common given names or parts of given names such as 'Emily' and 'Stephen', or obvious gender indicators such as 'girl' and 'woman'. 'Guy' and 'boy' occupy ranks 19 and 20 for male usernames, respectively. We assume that humans use similar cues to infer gender from usernames.

After this initial evaluation, the predictions of the classifier were compared to guesses made by humans as well as to the ground truth. 600 usernames were sampled from among the users whose real gender was known. Because the sampling method affects the distribution of categories in the sample, which in turn could influence the annotation results, half of the usernames were obtained by random sampling (resulting in 10% women), while the other half were obtained by stratified sampling (ensuring that both categories were equally frequent). Eight participants each received 100 randomly sampled usernames, and another eight participants each received 100 usernames obtained by stratified sampling. The participants were asked to indicate the likely gender of each user. They were given three options (man, woman, and unsure). If they were unsure, they were nevertheless requested to give their best guess. Participants were asked to rely on their gut feeling, and to spend ten seconds or less on each name. This combination of two questions allows us to gauge how certain users were of their guesses.

Ideally, the classifier should agree with a random human as often as two random humans agree with each other. When the 1600 guesses by the humans are compared with the classifier's predictions, they agreed in 76.0% of cases. In our data set, there were 1500 pairs of annotations, i.e. in 1500 cases, the same username was annotated by two different participants. In 75.8% of cases, both annotators agreed with each other. It appears from these figures that the classifier is about as good at determining gender based on usernames as human annotators.

The model also quantifies its certainty that a given username belongs to a man or woman. Ideally, the higher the probability of a user being a particular gender, according to the model, the more often humans should agree with this guess. Figure 1 shows that as the threshold of certainty required of the model increases, so does percent agreement with the human guess. When, for example, only those users are taken into account whose gender the model is 90% certain of, the human annotators agreed with this prediction 87.9% of the time. When the minimum certainty threshold is raised to 99%, agreement increases to 95.2%. These results show that not only are the model's predictions in line with human guesses, but the probabilities produced by the model are also in line with how likely humans are to agree.

For completeness, we report the accuracy of the human guesses compared to the ground truth, in Table 2. As can be seen, there is indeed a large difference between the random sample and the stratified sample. The results for the stratified sample show that the humans are clearly better than random guesses (which would result in a precision and recall of 50% on both categories), but the results for the random sample show that identifying the few women in the userbase by their names is hard for both the model and the humans.

Finally, we used the model trained on the Last.FM data to calculate predictions for the entire data set of 1,598,796 Wikipedia users who had commented at least once on a talk page by December 2017.

3 RESULTS

The method we developed allows us to report descriptive statistics about how many Wikipedia users choose names that contain obvious references to their gender. Figure 2 visualises the distribution of gender markers in the data set. Only 4,6% of users (72,673 out of 1,598,796) have chosen a name that is more female than male (i.e. the probability of it being female is larger than 50%). Only 1.2% of usernames are clearly female (probability > 90%). In contrast, 80.0% of usernames are clearly male.

Using the classifier, it is also possible to address the original research question: Are posts by users with female usernames less likely to receive a reply than posts by users with male names? To address this question, we exclude ambiguous names from the analysis, i.e. those with a probability between 10% and 90% of being female. We compared posts by the remaining users with obvious gender markers regarding their reply rate. The remaining users make up 81,2% of all Wikipedia users who wrote at least one comment. In addition, we only considered top-level posts in this analysis. By

Table 4: Reply probability by perceived gender

Top-level posts …	Number of posts
by authors with a male name	4,975,671
— which received at least one reply	1,667,783 (33,52%)
by authors with a female name	163,520
— which received at least one reply	57,110 (34.93%)

top-level posts, we refer to posts that begin a new discussion thread, as opposed to replies to existing posts. A reply to an existing post that does not, in turn, receive a reply, could simply indicate that the issue has been resolved or the question answered.

The results are presented in Table 4. As can be seen, usernames that were obviously female were not less likely to be replied to – in fact, they appear to be more likely to be replied to. The present data do not support the hypothesis.

4 DISCUSSION

In the process of addressing the hypothesis of this paper, we developed a classifier that determines whether a username contains obvious gender markers. The classifier was trained on 1.4 million internet usernames. It therefore goes beyond simply looking for given names or words such as 'woman' but contains a dictionary of markers that are particular to online usernames, such as 'grrl' and 'lord'. The classifier was validated by comparing it with the judgments of 16 humans. Agreement between our classifier's and human annotators' predictions is at 76% and agreement between human annotators is at 75.8%. Therefore, our classifier would be useful as a research tool to automatically examine gender markers in large data sets of usernames in a manner that mimics human judgments.

The present data do not support the hypothesis that there are differences in reply rate between users with masculine markers and users with feminine markers in their usernames. Therefore, it seems implausible that the differences in reply rate between men and women, which we showed in an earlier paper [2], are due to gender markers in their usernames. Instead, based on the present data, we can rule out this possibility. This result is important for the Wikipedia community because it implies that we found no evidence of discrimination of female users based on their usernames, unlike what other studies have found in offline and online correspondences in male-dominated fields [5, 7]. More research will be necessary to determine the causes of the statistically significant differences in the reply rate between men and women, especially in areas such as Engineering, where the difference is large. Other factors may be at play here, perhaps not all of which are measurable in observational data sets.

An analysis of the entire dump of all Wikipedia talk pages revealed that only 1.2% of users have obviously female names (probability >90%). Only 4.6% have names that appear more female than male (probability >50%), which is much lower than the number of female users in the data set of users whose gender we know (10.8%). It corroborates earlier results that some women avoid using gender markers in their usernames [6, 11]. This is especially surprising since our results also show that, as far as reply rate is concerned, users with obviously female names are not more likely to be ignored in talk pages. In fact, having more usernames with feminine markers may help to bridge the gender gap on Wikipedia, since previous research has shown that seeing active users with feminine-sounding usernames encourages participation among women [9]. Therefore, encouraging women to use feminine markers in their Wikipedia usernames may help to increase visibility and participation of women in the community, without having to worry about negative consequences that arise from displaying their gender identity. Further work could investigate whether contributions by women increase over time in articles that were proposed or predominantly edited by users with feminine usernames.

REFERENCES

[1] David Bamman and Noah A Smith. 2014. Unsupervised discovery of biographical structure from text. *Transactions of the Association for Computational Linguistics* 2 (2014), 363–376.

[2] Benjamin Cabrera, Björn Ross, Marielle Dado, and Maritta Heisel. 2018. The Gender Gap in Wikipedia Talk Pages. (2018). Submitted for publication.

[3] Benjamin Cabrera, Laura Steinert, and Björn Ross. 2017. GraWiTas: a Grammar-based Wikipedia Talk Page Parser. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. 21–24. https://doi.org/10.18653/v1/E17-3006

[4] Shyong (Tony) K. Lam, Anuradha Uduwage, Zhenhua Dong, Shilad Sen, David R. Musicant, Loren Terveen, and John Riedl. 2011. WP:Clubhouse?: An Exploration of Wikipedia's Gender Imbalance. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration, WikiSym '11*. ACM, New York, NY, USA, 1–10. https://doi.org/10.1145/2038558.2038560

[5] Katherine L Milkman, Modupe Akinola, and Dolly Chugh. 2015. What happens before? A field experiment exploring how pay and representation differentially shape bias on the pathway into organizations. *Journal of Applied Psychology* 100, 6 (2015), 1678.

[6] Jane Miller and Alan Durndell. 2004. Gender, language and computer-mediated communication. *WIT Transactions on Information and Communication Technologies* 31 (2004).

[7] Corinne A. Moss-Racusin, John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman. 2012. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences* 109, 41 (2012), 16474–16479. https://doi.org/10.1073/pnas.1211286109 arXiv:http://www.pnas.org/content/109/41/16474.full.pdf

[8] Joseph Reagle and Lauren Rhue. 2011. Gender bias in Wikipedia and Britannica. *International Journal of Communication* 5 (2011), 21.

[9] Christina Shane-Simpson and Kristen Gillespie-Lynch. 2017. Examining potential mechanisms underlying the Wikipedia gender gap through a collaborative editing task. *Computers in Human Behavior* 66 (2017), 312 – 328. https://doi.org/10.1016/j.chb.2016.09.043

[10] Kaveri Subrahmanyam, Patricia M. Greenfield, and Brendesha Tynes. 2004. Constructing sexuality and identity in an online teen chat room. *Journal of Applied Developmental Psychology* 25, 6 (2004), 651 – 666. https://doi.org/10.1016/j.appdev.2004.09.007 Developing Children, Developing Media - Research from Television to the Internet from the Children's Digital Media Center: A Special Issue Dedicated to the Memory of Rodney R. Cocking.

[11] Bogdan Vasilescu, Andrea Capiluppi, and Alexander Serebrenik. 2014. Gender, Representation and Online Participation: A Quantitative Study. *Interacting with Computers* 26, 5 (2014), 488–511. https://doi.org/10.1093/iwc/iwt047