Inferring Sociodemographic Attributes of Wikipedia Editors: State-of-the-art and Implications for Editor Privacy

Sebastian Brückner
RWTH Aachen University
Germany
sebastian.brueckner@rwth-aachen.de

Florian Lemmerich
RWTH Aachen University
Germany
florian.lemmerich@cssh.rwth-aachen.de

Markus Strohmaier
RWTH Aachen University &
GESIS - Leibniz Institute for the Social Sciences
Germany
markus.strohmaier@cssh.rwth-aachen.de

ABSTRACT

In this paper, we investigate the state-of-the-art of machine learning models to infer sociodemographic attributes of Wikipedia editors based on their public profile pages and corresponding implications for editor privacy. To build models for inferring sociodemographic attributes, ground truth labels are obtained via different strategies, using publicly disclosed information from editor profile pages. Different embedding techniques are used to derive features from editors' profile texts. In comparative evaluations of different machine learning models, we show that the highest prediction accuracy can be obtained for the attribute gender, with precision values of 82% to 91% for women and men respectively, as well as an averaged F1-score of 0.78. For other attributes like age group, education, and religion, the utilized classifiers exhibit F1-scores in the range of 0.32 to 0.74, depending on the model class. By merely using publicly disclosed information of Wikipedia editors, we highlight issues surrounding editor privacy on Wikipedia and discuss ways to mitigate this problem. We believe our work can help start a conversation about carefully weighing the potential benefits and harms that come with the existence of information-rich, pre-labeled profile pages of Wikipedia editors.

CCS CONCEPTS

• **Human-centered computing → Wikis**.

KEYWORDS

Wikipedia, editor attribute prediction, BERT

ACM Reference Format:
Sebastian Brückner, Florian Lemmerich, and Markus Strohmaier. 2021. Inferring Sociodemographic Attributes of Wikipedia Editors: State-of-the-art and Implications for Editor Privacy. In *Companion Proceedings of the Web Conference 2021 (WWW '21 Companion), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3442442.3452350

1 INTRODUCTION

Wikipedia is one of the most frequently visited websites on the internet worldwide [30]. For many people, Wikipedia represents a key source of information for a wide variety of purposes [13, 26]. Due to its importance to a global audience, understanding the sociodemographic composition of the community that is producing the content of Wikipedia is an essential prerequisite for tackling current challenges related to Wikipedia, such as topical or social biases, inclusiveness and discrimination. At the same time, little is known about the sociodemographic composition of Wikipedia's editors. An early survey conducted in 2011 [7] found that among those editors who have answered questions, 91% were male and only 8.5% were female. Since then, similar surveys have been repeated, finding that roughly 85% of the users stated being male [29]. While the execution of Wikipedia editor surveys is costly and can not be easily done with adequate coverage or high temporal resolution, automatic approaches towards inferring sociodemographic attributes of Wikipedia editors promise to overcome these challenges, but exhibit a clear potential for producing unintended side effects such as compromising editor privacy.

Objective. As a consequence, this paper aims to i) evaluate the feasibility of different machine learning models to predict sociodemographic attributes of Wikipedia editors based on their public profiles, and to ii) increase awareness for resulting concerns related to the privacy of Wikipedia editors.

Approach. For training and evaluation of different machine learning classifiers, we collect the publicly accessible profile pages of registered editors of the English language Wikipedia edition. These include both texts written by the editors about themselves, which are used as input features, as well as so-called user boxes displaying personality traits, which are used as labels. The latter is complemented by information about categories a user chose to associate with, e.g., *Category: Female Wikipedians* and used as a ground truth for sociodemographic attributes. After performing feature embedding on the profile texts, we assess the performance of several classification algorithms on sociodemographic attribute inference.

Contributions. We present a comparison of different machine learning models for inferring sociodemographic attributes of Wikipedia editors, based on Wikipedia's user profile pages. We find significant differences in the predictive power of these models regarding different sociodemographic attributes: While gender can be predicted with the highest accuracy, other attributes perform worse. Inferring sociodemographic attributes of Wikipedia editors could enable studies of the underlying bias in the distribution of Wikipedia editor demographics, differences in behavior, perceived experience, and influence on the platform. This could potentially be used to analyze controversial articles for bias caused by different

WWW '21 Companion, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8313-4/21/04.
https://doi.org/10.1145/3442442.3452350

editor demographics, e.g., it would be interesting to know whether or not an article about *abortion* was predominantly written by men or women.

At the same time, our analysis highlights fundamental limitations of attribute inference for Wikipedia editors. We urge caution with respect to the deployment of such inference models: The prediction of sensitive attributes should not only be carefully assessed with respect to ethical considerations, but one should also be aware of the limits of the prediction of sociodemographic attributes on an individual level. We believe that while sociodemographic attribute inference models could potentially add useful information to the analysis of social dynamics and issues on Wikipedia, they should only be considered to derive global statistics on an aggregated (vs. individual) level.

Implications. Our work informs the Wikipedia editor community that their publicly provided profile data could not only be used by researchers, but potentially also by others with malicious intent, e.g., for profiling purposes. Moreover, revealing information about themselves can enable the training of machine learning models for inferring sociodemographic attributes of *other* editors who might have chosen *not to reveal* this information. As a consequence, we recommend the Wikipedia editor community to carefully weigh the potential benefits and harms of providing detailed information about themselves on editor profile pages.

2 RELATED WORK

Our work relates to different areas of previous research on Wikipedia and collaborative authoring, which we will outline in this section.

Wikipedia and research on sociodemographics. Rizoiu et al. [22] investigate the digital tracks left by Wikipedia users to discover some of their undisclosed private traits. Their approach is solely based on the publicly available feature of the number of contributions within a specific set of categories for a given editor. The main focus is on how the prediction accuracy for labels like gender, education, and religion improves over the analyzed time span of edits. The authors observe that the longer the considered editing history is, the more accurate the predictions become. The authors refer to this observation as the *evolution of privacy loss*. For obtaining a ground truth, the information provided by users on their profile pages is collected using the MediaWiki API [14]. Another approach for age and gender prediction of *blog authors* by Santosh et al. [24] made use of Wikipedia by extracting entity mentions in the text to analyze and collect the corresponding Wikipedia concepts and related categories. According to the authors, this is supposed to help with polysemy and overall accuracy. Bérubé et al. [3] developed a first name-based gender detection algorithm using Wikipedia. It works by mining content from articles about people and subsequently associating each person with a gender via analyzing keywords contained in that text employing various methods. This results in a first-name-to-gender mapping table. In an effort to determine if the reason for female editors receiving fewer replies than their male counterparts are gender clues in their usernames, Ross et al. [23] developed an algorithm for determining gender markers in Wikipedia usernames. The authors found that editors with clearly identifiable female names receive more replies; however, men are more likely to have a username that hints at their gender. Karimi et al. [9] evaluate different ways to infer gender from names on the web. A taxonomy of knowledge gaps for Wikipedia projects has been published in [21].

Author profiling. Our work relates to research on author profiling. In this direction, Schwarz et al. [25] utilized texts from Facebook messages together with personality tests to find relationships between linguistic styles and attributes like gender and age as well as personality traits. The authors used differential language analysis in an open-vocabular approach to obtain linguistic features that describe differences in demographic attributes. As a result, they found that females use words related to emotions and first-person pronouns more frequently, while males tend to utilize more swear words. Furthermore, an SVM machine learning model was trained for automatic gender prediction. As part of the PAN at CLEF initiative [1] there have been multiple author profiling challenges in the past years, which generated a substantial body of research in this field, summarized in [18], [20] and [19].

Wikipedia talk pages. There is substantial research on Wikipedia's talk pages, where editors can communicate with each other about issues of the corresponding article. For example, Cabrera et al. [4] studied gender bias by gathering a comment dataset together with editors' gender information extracted from their profile pages. The authors show that overall, men are more likely to write comments, especially in male-dominated fields like engineering or physics. Additionally, females have a statistically significant lower probability of receiving a reply to their comment, with men and women being more inclined to reply to their own gender, respectively.

As talk pages can be the place for heated discussions among editors, they are also prone to harassment, insults, and toxicity. In order to address, moderate and visualize this problem, Qu et al. [16] present the *WikiDetox* tool, leading to the development of *Perspective API* [8]. The utilized machine learning models evaluate user comments with respect to general aspects like toxicity and more nuanced issues like flirtation and sexual explicitness. Data from Wikipedia dumps is then analyzed using the proposed methods, resulting in a visualization of interactive word clouds showing toxic and detoxed comments. The latter refers to deleted texts.

In another paper focusing on personal attacks on Wikipedia by Wulczyn et al. [31], a machine learning classifier is trained on human-labeled data to facilitate the identification of problematic comments within a broader scope. The authors conclude that a comment made by anonymous users is six times more prevalent to constitute an attack. Nevertheless, also highly active editors, with more than 100 contributions, clash with each other, making up roughly 30% of all attacks.

Wikipedia gender bias. Furthermore, a substantial amount of research has been conducted with respect to Wikipedia's gender bias. For example, Antin et al. [2] examine the gender gap by looking at the differences between what women and men tend to do on Wikipedia. The authors state that contrary to the skewed distribution of editors' gender, male and female users make similar numbers of revisions when only considering the lower 75% of editors sorted by activity level, who are responsible for roughly 9% of all revisions. Additionally, the most engaged female users make even more extensive edits than their male counterparts. Nevertheless, men contribute significantly more edits when restricting the view to

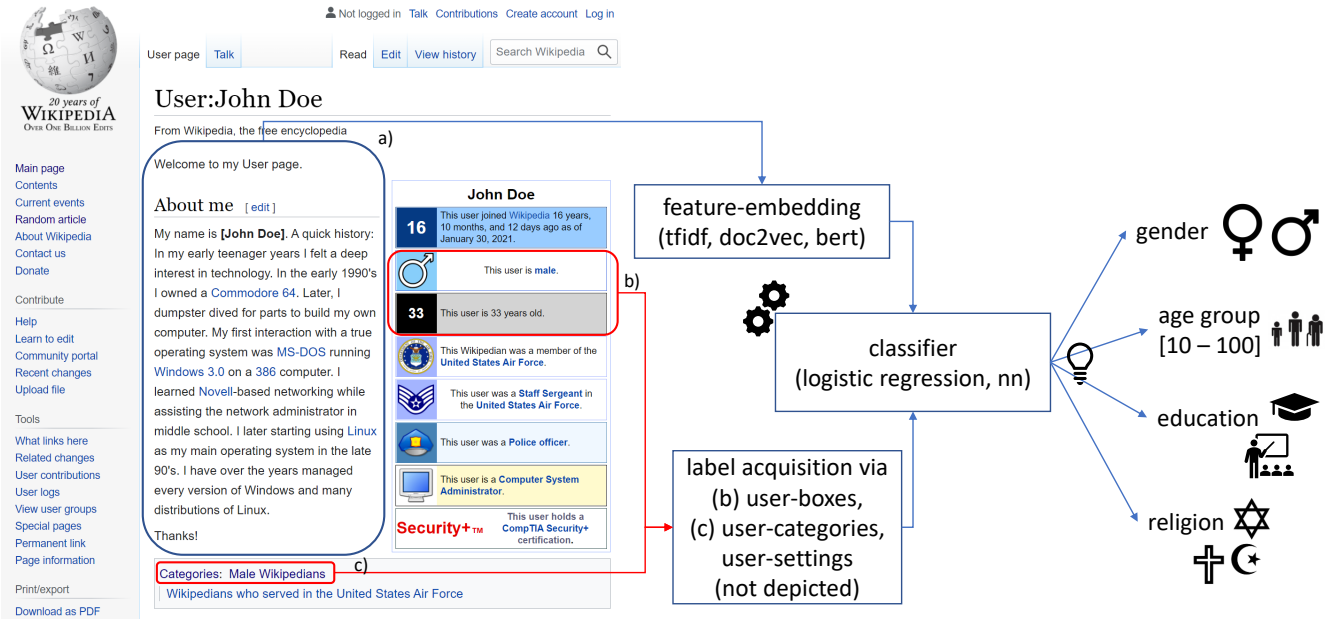

Figure 1: Approach. Different embedding methods are used to extract features from (a) the profile texts of Wikipedia editors which are given as input to the classifier. The corresponding labels are extracted from (b) user-boxes, (c) user-categories or from publicly accessible user-settings (not depicted). A classifier is used to predict the sociodemographic attributes gender, age group, education, and religion.

the most active users. Gender bias was also analyzed by Lam et al. [11], revealing differences in work performed in various areas of content, with males participating more in science while women focus more on arts. They additionally discover that the first few edits of an account operated by a woman have a significantly higher probability of being reverted than similar revisions contributed by male editors. It is also determined that articles with higher than average female editor participation are more prone to be controversial and that the likelihood for female users to be permanently blocked is higher. Wagner et al. [28] have studied gender bias of notable people on Wikipedia by investigating textual, metadata and hyperlink features. The authors report significant differences in the language used to describe notable people, and the way notable people are linked within the Wikipedia hypertext network.

3 APPROACH

To acquire training data, we collect a list of usernames of Wikipedia editors, the text on each user's user page, as well as a corresponding label for considered sociodemographic attributes. The former can be obtained by crawling Wikipedia's special page *Users*. The profile text of every given user can be obtained by using the query module of the MediaWiki API [14]. It provides means for stating queries such that different types of data pages, properties of these pages, and metadata information can be requested. Lastly, the labels for the different author attributes can be collected from several sources: a user's publicly accessible profile settings, membership in attribute revealing categories as well as usage of so-called userboxes. The latter represents a form of badges used by editors to display anything from their achievements to very specific personality traits, e.g., hobbies or which browser they use to access Wikipedia. This also includes the attributes gender, age, religion, and level of education. An overview of our approach, visualizing the interaction of feature embedding, label acquisition, and the classifiers used, is provided in figure 1.

3.1 User Profile Text Acquisition

According to Wikipedia's special page *Users*, at the time of data acquisition, the English Wikipedia had 39,322,024 registered users. However, only 11,335,000 of those had contributed at least one edit. Furthermore, having a profile page is a necessary requirement for our experiments since a user's profile text is used as the input feature for the classifier. This constraint reduces the number of relevant usernames to 1,742,506.

In the next step, the profile texts of all editors whose name was included in the initially crawled username list is obtained using the MediaWiki API. Of the 1,742,506 queried user pages, 8,559 did not return a result. The majority of those pages just linked to a corresponding user page in the meta-wiki, which were crawled again, this time with an updated URL. This worked for all but 1,130 users due to one of the following reasons: the account was deleted, the account was banned, or the user was renamed.

3.2 Sociodemographic Label Acquisition

We derived labels for sociodemographic attributes of Wikipedia editors as ground truth for training a machine learning classifier in the following way:

Gender. The gender labels of editors were gathered using the three aforementioned approaches. First, users' affiliation to gender revealing categories, e.g., *Male Wikipedians* or *Female Wikipedians*, is used. This yields 15,012 male and 3,561 female labels. Secondly, the template usage in the form of user boxes placed on editors' pages is examined. Similar to the mentioned categories, certain user boxes indicate a person's gender, e.g., *This user is a woman*. Using this as a basis for extracting labels results in 9,996 male and 1,783 female labels. However, only 1,648 male and 288 female editors were previously not known. Lastly, the list of usernames obtained in the first step was used to query the MediaWiki API for each user's corresponding gender. The API only returns a result if the user in question chose to disclose his or her gender. This way, 109,063 new male and 19,361 new female users could be gathered. In order to test the quality of the three different sources, they were checked against each other, yielding virtually no contradictions. In total, of the 1,742,506 users that were queried, 148,933 disclosed their gender, which is roughly 8.5%. Of those, 125,723 are male, and 23,210 are female, resulting in a percentage distribution where 84% are male, and 16% are female. This finding is fully concordant with the findings of the Wikipedia survey of 2011 [7].

Age. For age prediction, the same features, i.e., profile texts, are used as for gender prediction, where 2,583 users disclosed their age. The corresponding labels were obtained by extracting information from categories the user belongs to, as well as user boxes stating "This user is X years old" placed on the user pages. Afterward, the extracted age values were sanity checked, removing all labels smaller than 10 and greater than 100 as some users stated clearly invalid values such as an age of 1,000 years. Furthermore, additional data [10], in the form of blog posts, outside of Wikipedia was used for training the classifier due to the small number of available labels. Said data consists of textual features that are similar to editor's profile page posts and corresponding age labels. Finally, the labels were subdivided into the three following classes: <18, 18-30, >30.

Education. The prediction of an editor's education follows the same principle as with the previously mentioned attributes. Profile-texts are used as input features, while the labels are obtained from categories associated with the users as well as unambiguous user boxes. The 11,142 obtained labels include users who stated that they did *not have a high-school education*, are *school students*, are *undergraduates*, are *graduates*, or have a *Ph.D.* For predictions, only the last three were chosen due to the very small amount of labels for the first two. Similar to the attribute gender, a noticeable label imbalance for education labels is present, with more than twice as many labels corresponding to *undergraduates* as compared to either one of the two other classes. Furthermore, if users had two different labels regarding their education, only the label referring to the highest level of education is kept.

Religion. Editors' religion was predicted based on 8,746 queried labels for the classes *Christianity*, *Atheism*, *Islam*, and *Judaism*. The labels are not evenly distributed, with *Christianity* and *Atheism* both having more than twice as many labels as either one of *Islam* and *Judaism*.

In a final step, the gathered text and label information were merged. This revealed that for 103 users with known labels no profile text was queried initially. Possible reasons for this could be that the users in question did not perform any edits and were therefore not included in the username list, that they do not have a profile page, that their account was deleted, or that their page is part of the meta-wiki. In the latter case, the profile text was additionally crawled.

3.3 Feature engineering

To use the crawled texts as input for a machine learning classifier, we need to embed the strings into numerical representations (features). Next, we briefly outline the different techniques we employed for this task.

Tf-idf. Tf-idf is based on the word frequencies [17]. First, tokenization is used to split up the text into individual items, so-called tokens. Those words that frequently appear in any text and therefore carry very little meaning, e.g., articles like *the* and *a* are removed, reducing the size and complexity of the token list. Nevertheless, the updated collection of tokens still needs further processing since it inherently suffers from a high variation of natural language. The problem is that a tokenized verb could have been collected in various conjugations, increasing the complexity of the data without providing additional useful information. To address this, different variations of a word are truncated based on their word stem. For example, the occurrences of *talk*, *talking* and *talked* would simply be changed to *talk*.

Afterwards, the transformation from text into numerical values is achieved by counting the number of occurrences of words broken down by documents. Words with very high numbers of appearances often do not carry significant informational value; their effect on calculations can, however, be very noticeable. To combat this, a method from *scikit-learn* is used, namely a *Tf-idf-Transformer*, which mitigates their influence on the outcome, by substituting the word frequencies by a *Tf-idf*-score. The score is given by the following formula:

$$\text{tf-idf(t,d)} = tf(t,d) * idf(t)$$

Here, d stands for an arbitrary document, t for an arbitrary term and *idf(t)* for its inverse document frequency. The latter is determined as follows in the *scikit-learn* library [15]:

$$idf(t) = log\left(\frac{1+n}{1+df(d,t)}\right) + 1$$

In the above-mentioned equation, n describes the total number of documents and *df(d,t)* denotes the number of documents that contain term t. To avoid terms with occurrences in every document to have a score of zero, the one is added at the end, providing them with at least some amount of importance [15].

Doc2Vec. The second approach is called Doc2Vec [12]. It can be used to directly calculate individual document embeddings while simultaneously overcoming two major shortcomings of the previous approach, which neglects the word ordering and word semantics. Doc2Vec follows a similar approach to the more popular Word2Vec, where word vectors are used to predict the next words in a given context, which results in obtaining semantics [12]. Comparably,

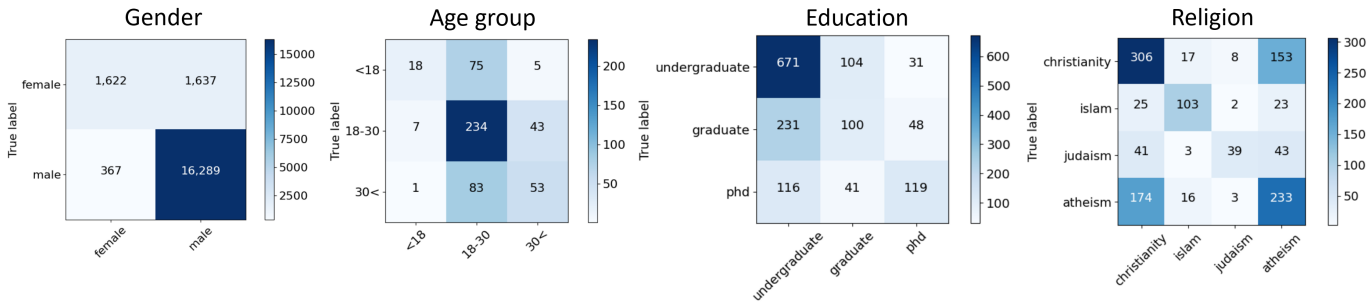

Figure 2: Confusion matrices for gender (90% accuracy), age group (59% accuracy), education (61% accuracy), and religion (57% accuracy) prediction based on BERT. The matrices corresponding to the various attributes clearly show high label imbalance and the related misclassification rate. For example, more than half of all female labels are incorrectly classified as male user. Overall, in terms of precision, the attribute gender can be predicted best.

the document vectors in Doc2Vec are also used to predict the next word in a context, based on an excerpt of a given document. The mappings of words and documents to vectors are stored in the matrices, which are learned during training as a side product of the next-word prediction task [12].

BERT. The third approach is a language representation model called *BERT* [6]. It can basically be understood as a pre-trained representation of text, in the form of a neural network, that can be fine-tuned depending on the problem at hand by merely adding an output layer. The original implementation of the *multi-layer bidirectional Transformer encoder* by Vaswani et al. [27] serves as architecture for the BERT model.

An important property of BERT is the fact that each word has its own dedicated path through the encoder and the only dependencies between words are present in self-attention layers. This so-called attention mechanism allows the model to see relationships between individual words by giving words different weights based on the context. For example, given the sentence "The man was tired, but he did not know why", attention would enable the model to know what the word *he* refers to.

BERT's distinctive feature is that its bidirectional representations are pre-trained based on the right and left context. This is achieved by utilizing a masked language model, which is comprised of providing the model with a piece of unlabelled text, in which certain words are masked at random. This pursues the goal of using the given context to predict the masked word. In order to perform fine-tuning, the input words have to be specifically preprocessed for BERT. First, the model has a fixed input length, i.e., any text longer than that has to be shortened accordingly. Next, each word is tokenized, creating a list of tokens. At the first position of the list, a special *[CLS]* token is placed that contains the hidden state of the model if it is used for a classification task. The list is completed by a *[SEP]* token that indicates the end of the text. Lastly, each token is converted to a numerical id according to the vocabulary list provided with the pre-trained BERT model.

3.4 Prediction of sociodemographic attributes

Overall, the three methods Tf-idf, Doc2Vec, and BERT were used in combination with a diverse set of classifiers including Logistic Regression (LR), Stochastic Gradient Descent (SGD), and Feed-forward neural networks (NN), see Table 1. With the exception of neural networks, we used implementations from *scikit-learn* [15]. For the NN, we used *TensorFlow* and experimented with multiple configurations. The best performances were achieved using the pretrained BERT model, followed by a dropout layer set to 0.5, a dense layer with 768 nodes, another dropout layer with the same parameters, and a subsequent softmax layer.

Since the class distributions for several sociodemographic attributes are highly skewed, e.g., in the training data, there are many more male editors compared to female editors, we experimented with several re-sampling methods like SMOTE [5] to oversample the minority class and undersample the majority class, which did not improve the performance. In order to use the best parameters for logistic regression, grid-search via 5-fold cross-validation with the F1-score as optimization score was used. However, applying the resulting optimized parameters yields a slightly better recall for the minority classes at the expense of even worse precision. Therefore, the standard parameters were utilized.

For both the Doc2Vec and Tf-idf embedding, the same tokenized and stemmed data was used as input. By contrast, BERT's input features are based on the aforementioned word to id mappings based on a vocabulary list.

4 PREDICTION PERFORMANCE

Training and subsequent evaluation were implemented by splitting the data into separate training (80%) and test (20%) sets. The classification results are listed in table 1, with confusion matrices based on the best performing approach (BERT) shown in figure 2.

Gender. When comparing the scores for the different types of embeddings for the attribute gender, Doc2Vec was found to be inferior compared to Tf-idf and BERT, with Tf-idf having a high precision but low recall for female labels. In contrast, BERT has both a high precision as well as a noticeable higher recall for females. One major issue is that the classifiers have a high incentive to favor male predictions, which is caused by class imbalance. Classifying every user as male would already yield a high accuracy on average. Nevertheless, using BERT still results in high precision for female users, at the expense of recall. As can be seen in the confusion matrix for gender in figure 2, more than half of all females are wrongly classified as males, which does not affect male precision

Table 1: Prediction results (precision, recall, and F1-score) for different models. The results show that BERT outperforms Tf-idf and Doc2Vec for all sociodemographic attributes, and predictions for gender are more accurate than the predictions for all other attributes.

	Tf-idf - LR			Tf-idf - SGD			Doc2Vec - LR			BERT - NN		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Male	0.87	0.99	0.93	0.85	0.99	0.91	0.86	0.98	0.92	0.91	0.98	0.94
Female	0.82	0.26	0.39	0.89	0.12	0.21	0.64	0.16	0.26	0.82	0.5	0.62
<18	0	0	0	0.38	0.21	0.27	0.22	0.04	0.06	0.56	0.29	0.38
18-30	0.55	0.91	0.68	0.56	0.66	0.6	0.57	0.9	0.69	0.58	0.77	0.66
30<	0.36	0.1	0.15	0.35	0.33	0.34	0.47	0.17	0.25	0.49	0.34	0.4
undergraduate	0.63	0.92	0.75	0.65	0.82	0.72	0.6	0.89	0.72	0.66	0.83	0.74
graduate	0.41	0.17	0.24	0.35	0.24	0.28	0.36	0.12	0.18	0.41	0.26	0.32
Phd	0.64	0.25	0.36	0.51	0.33	0.4	0.48	0.21	0.29	0.6	0.43	0.5
Christianity	0.5	0.65	0.56	0.53	0.59	0.56	0.5	0.54	0.52	0.56	0.63	0.59
Atheism	0.47	0.53	0.5	0.49	0.54	0.51	0.45	0.53	0.49	0.52	0.55	0.53
Islam	0.9	0.38	054	0.78	0.65	0.71	0.49	0.46	0.48	0.74	0.67	0.71
Judaism	1	0.09	0.16	0.59	0.26	0.36	0.44	0.11	0.18	0.75	0.31	0.44

much, simply because there are many more males. If the classifier predicts a user to be female, this is correct in 82% of the tested cases. To sanity-check the distribution of predictions for the attribute gender, we applied the BERT model, with a confidence threshold of 95%, to all user pages that are not included in the ground-truth. 89.74% of those users were determined to be men, and 10.26% to be women, compared to 84% and 16% in the ground truth, respectively.

Age. Predictions regarding age are considerably less accurate compared to gender. The best performance is achieved again by BERT with precision values for all classes around 50% and low recall values of 29% and 30% for the classes *<18* and *30<*, respectively. However, interestingly, users from either one of the age groups that are farthest apart, i.e., *<18* and *30<*, are only very rarely mistaken for each other as can be seen in the corresponding confusion matrix.

Education. Regarding the prediction of education, Doc2Vec in combination with LR has the worst performance, while Tf-idf and BERT perform quite similarly. Nevertheless, BERT still achieves higher recall values for the two classes *graduate* and *Phd* resulting in a higher F1-score on average. Looking at the confusion matrix for education, it becomes clear that the classifier seems to favor predicting users as belonging to the class *undergraduate*, which is probably caused by the aforementioned imbalance.

Religion. Lastly, the results for the attribute religion are in line with the other findings, with Tf-idf outperforming Doc2Vec but BERT still being superior overall. The latter achieves precision scores in the range of 50% - 70% with similar recall values. The average F1-score is 0.57.

When looking at the corresponding confusion matrix, it seems like most misclassifications are caused by Christians being mistakenly predicted to be atheists and vice versa.

5 LIMITATIONS AND IMPLICATIONS

We discuss general challenges and limitations of sociodemographic attribute inference from user profiles on Wikipedia, and consequences for editor privacy.

Data limitations. On a fundamental level, the predictive capabilities of the models studied here are limited by the amount of text that editors write about themselves on their profile page. If there is no text, no reasonable prediction can be obtained. This ties in a more wide-spread elementary concern: editors that explicitly mention sociodemographic attributes on their profile pages might not be representative (with respect to their overall profile page content) of the respective sociodemographic group of editors. This means that any potential biases underlying the production of text on profile pages will manifest in the resulting prediction models. Furthermore, our evaluation of prediction approaches assumes that users who disclose their sociodemographic attributes do so accurately, or at least that there is no significant difference in incorrectly posted attributes for the various classes. It is possible that this is not the case, and/or that a disclosure bias is present, i.e., that those who willingly share their personal information are different than the entire population. This would inject significant biases in the prediction results. Furthermore, the labels used for training might be reductionist, e.g. using binary labels for the gender attribute. Thus, the performance scores presented in our evaluation are likely more optimistic compared to what one could expect for extrapolating the classifier to arbitrary editors.

Implications for editor privacy. In our experiments, we investigate the extent to which sociodemographic attributes can be inferred from user profiles on Wikipedia. We emphasize that such analysis has to be conducted with extreme care. For example, we recommend to only utilize the acquired labels on an aggregated level to find general disparities in Wikipedia's editor community or in selected, large groups of articles. To use this kind of information on an individual level, e.g., for personalized recommendation, effects editor privacy and might result in unethical applications. This could lead to frequent misattributions considering the reported overall classification performance that was achieved.

Finally, and perhaps most importantly, our work informs Wikipedia editors that their publicly provided profile data could not only be used by researchers, but potentially also by others with malicious intent, e.g., for profiling purposes. Alleviating these potential harmful consequences however requires more than just individual editors stopping to reveal sociodemographic information about

themselves. A large enough set of editors deciding to disclose sociodemographic attributes would still enable predictions for editors *who may have decided against it*. For more comprehensive guarantees of privacy, the editor community might want to critically assess the use of profile pages *at all*. Less drastic steps could involve *not* utilizing user boxes, thereby making automated processing and predictions more difficult, since they can be directly used as class labels during training. Another counter measure could be to aim for reducing the quality of one's input features, e.g., by reducing the amount of text written one's user page.

6 CONCLUSIONS

This paper aimed to evaluate the feasibility of different machine learning models to predict sociodemographic attributes of Wikipedia editors and to increase awareness for concerns related to the privacy of Wikipedia editors by highlighting and discussing some of the implications of editors who are making personal information public via profile pages. We report results from a comparative evaluation of several machine learning models. Our results show that across different models, (binary) gender can be predicted best, having high precision values and an averaged F1-score of 0.78. Regarding other attributes, i.e., age group, education, and religion, classification results exhibit significantly lower F1-scores, ranging from 0.32 to 0.74.

While prediction models could yield an interesting alternative to conducting costly editor surveys, our findings also suggest that caution is warranted given that across all attributes and approaches, biases can easily manifest and misclassifications are produced. In addition, establishing a robust ground truth for training still represents a significant challenge. In this paper, only heuristic approaches were pursued, future work might investigate the possibility of more elaborate approaches such as linking profiles with surveys. Overall, caution should be exercised when deploying sociodemographic inference models to Wikipedia profile pages. Any analysis following sociodemographic attribute inference should be conducted on an aggregated level only, and ethical issues should be taken into account.

REFERENCES

[1] 2020. PAN. https://pan.webis.de/. (accessed August 11, 2020).

[2] Judd Antin, Raymond Yee, Coye Cheshire, and Oded Nov. 2011. Gender differences in Wikipedia editing. In *Proceedings of the 7th international symposium on wikis and open collaboration*. 11–14. https://dl.acm.org/doi/pdf/10.1145/2038558.2038561

[3] Nicolas Bérubé, Gita Ghiasi, Maxime Sainte-Marie, and Vincent Larivière. 2020. Wiki-Gendersort: Automatic gender detection using first names in Wikipedia. (2020). https://osf.io/preprints/socarxiv/ezw7p/download

[4] Benjamin Cabrera, Björn Ross, Marielle Dado, and Maritta Heisel. 2018. The Gender Gap in Wikipedia Talk Pages. In *Twelfth International AAAI Conference on Web and Social Media*.

[5] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[7] Wikimedia Foundation. 2011. Wikipedia editors study: Results from The Editor Survey, April 2011. (2011). https://upload.wikimedia.org/wikipedia/commons/7/76/Editor_Survey_Report_-_April_2011.pdf

[8] Google and Jigsaw. 2020. Perspective API. https://www.perspectiveapi.com. (accessed August 11, 2020).

[9] Fariba Karimi, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi, and Markus Strohmaier. 2016. Inferring gender from names on the web: A comparative evaluation of gender detection methods. In *Proceedings of the 25th International conference companion on World Wide Web*. 53–54.

[10] Moshe Koppel, Jonathan Schler, Shlomo Argamon, and James Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI 2006 spring symposium on computational approaches to analysing weblogs*. 1–7.

[11] 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*. 1–10.

[12] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*. 1188–1196.

[13] Florian Lemmerich, Diego Sáez-Trumper, Robert West, and Leila Zia. 2019. Why the world reads Wikipedia: Beyond English speakers. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 618–626.

[14] MediaWiki. 2020. Manual:FAQ — MediaWiki, The Free Wiki Engine. https://www.mediawiki.org/w/index.php?title=Manual:FAQ&oldid=3888037 [Online; accessed 14-July-2020].

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[16] Iris Qu, Nithum Thain, and Yiqing Hua. 2019. WikiDetox Visualization. (2019).

[17] Anand Rajaraman and Jeffrey David Ullman. 2011. Mining of massive datasets: Data mining (ch01). *Min. Massive Datasets* 18 (2011), 114–142.

[18] Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. 2013. Overview of the author profiling task at PAN 2013. In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*. CELCT, 352–365.

[19] Francisco Rangel, Paolo Rosso, Manuel Montes-y Gómez, Martin Potthast, and Benno Stein. 2018. Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. *Working Notes Papers of the CLEF* (2018).

[20] Francisco Rangel, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. 2015. Overview of the 3rd Author Profiling Task at PAN 2015. In *CLEF*. sn, 2015.

[21] Miriam Redi, Martin Gerlach, Isaac Johnson, Jonathan Morgan, and Leila Zia. 2020. A Taxonomy of Knowledge Gaps for Wikimedia Projects (First Draft). *arXiv preprint arXiv:2008.12314* (2020).

[22] Marian-Andrei Rizoiu, Lexing Xie, Tiberio Caetano, and Manuel Cebrian. 2016. Evolution of privacy loss in Wikipedia. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. 215–224. https://dl.acm.org/doi/pdf/10.1145/2835776.2835798

[23] Björn Ross, Marielle Dado, Maritta Heisel, and Benjamin Cabrera. 2018. Gender markers in wikipedia usernames. In *Wiki Workshop*.

[24] K Santosh, Aditya Joshi, Manish Gupta, and Vasudeva Varma. 2014. Exploiting Wikipedia Categorization for Predicting Age and Gender of Blog Authors.. In *UMAP Workshops*.

[25] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one* 8, 9 (2013), e73791.

[26] Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. 2017. Why we read Wikipedia. In *Proceedings of the 26th international conference on world wide web*. 1591–1600.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).

[28] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 9.

[29] Wikipedia. [n.d.]. Community Insights/2018 Report. https://meta.wikimedia.org/wiki/Community_Insights/2018_Report. 2018 (accessed December 1, 2020).

[30] Wikipedia. 2020. Liste der meistaufgerufenen Websites. https://de.wikipedia.org/wiki/Liste_der_meistaufgerufenen_Websites. September 2018 (accessed July 12, 2020).

[31] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*. 1391–1399. https://dl.acm.org/doi/pdf/10.1145/3038912.3052591