

Towards Open-Domain Twitter User Profile Inference

Haoyang Wen^{*†}, Zhenxin Xiao^{*†}, Eduard H. Hovy^{†‡}, Alexander G. Hauptmann[†]

[†]Language Technologies Institute, Carnegie Mellon University

[‡]School of Computing and Information Systems, The University of Melbourne

{hwen3, zhenxin, hovy, alex}@cs.cmu.edu

Abstract

Twitter user profile inference utilizes information from Twitter to predict user attributes (*e.g.*, occupation, location), which is controversial because of its usefulness for downstream applications and its potential to reveal users' privacy. Therefore, it is important for researchers to determine the extent of profiling in a safe environment to facilitate proper use and make the public aware of the potential risks. Contrary to existing approaches on limited attributes, we explore open-domain Twitter user profile inference. We conduct a case study where we collect publicly available WikiData public figure profiles and use diverse WikiData predicates for profile inference. After removing sensitive attributes, our data contains over 150K public figure profiles from WikiData, over 50 different attribute predicates, and over 700K attribute values. We further propose a prompt-based generation method, which can infer values that are implicitly mentioned in the Twitter information. Experimental results show that the generation-based approach can infer more comprehensive user profiles than baseline extraction-based methods, but limitations still remain to be applied for real-world use. We also enclose a detailed ethical statement for our data, potential benefits and risks from this work, and our efforts to mitigate the risks. ¹

1 Introduction

Users' profile information provides invaluable user features. Accurate automatic user profile inference is helpful for downstream applications such as personalized search (Shen et al., 2005; Teevan et al., 2009; Zhu et al., 2008; Yao et al., 2020) and recommendations (Lu et al., 2015; Balog et al., 2019;

Guy, 2015), and computational social media analysis (Arunachalam and Sarkar, 2013; Bamman et al., 2014; Tang et al., 2015; Amplayo, 2019). However, there are increasing privacy concerns that conducting profiling without appropriate regulations may reveal people's private information. Therefore, it is essential to investigate the extent of profiling to promote proper use and make the potential risks clear to public and policy makers.


Previous work on user profile inference has focused on a very limited set of attributes, and models for different attributes employ different strategies. One line of research has formulated it as a classification problem for attributes such as gender (Rao et al., 2011; Liu et al., 2012; Liu and Ruths, 2013; Sakaki et al., 2014), age (Rosenthal and McKeown, 2011; Sap et al., 2014; Chen et al., 2015; Fang et al., 2015; Kim et al., 2017), and political polarity (Rao et al., 2010; Al Zamal et al., 2012; Demszky et al., 2019). In such classification settings, each attribute has its own ontology or label set, which is difficult to generalize to other attributes, especially for attributes that have many possible candidate values (*e.g.* geo-location, occupation). In addition, some work involves human annotation, which is expensive to be acquired and may raise fairness questions for labeled individuals (Larson, 2017).

Another line of research uses an extraction-based method, such as graph-based (Qian et al., 2017) and unsupervised inference (Huang et al., 2016) for geolocation, distant supervision-based extraction (Li et al., 2014; Qian et al., 2019). However, they still only cover limited attributes that cannot produce comprehensive profiles. Besides, many attribute values are only implicitly mentioned in Twitter context, which cannot be directly extracted.

In this paper, instead of limited attributes, we explore whether open-domain profiles can be effectively inferred. Taking WikiData (Vrandečić and Krötzsch, 2014) as the source of profile information, which provides a much more diverse

^{*}Equal contribution.

¹To prevent potential misuse of our resources, we will only release our resources for research purposes based on individual requests. Please see our ethics statement section for details and contact us directly for resource access.



Predicate	Value
Entity ID	Q76
Name	Barack Obama
Country of citizenship	United States of America
Occupation	Politician
Position held	President of the United States
Work location	Washington, D.C.
Spouse	Michelle Obama
...	...

(a) WikiData information.



Barack Obama
Dad, husband, **President**, citizen. **Washington, DC**

Across the country, **Americans** are standing up for abortion rights—and I’m proud of everyone making their voices heard. Join a march near you:

...

Happy Mother’s Day! I hope you all let the moms and mother-figures in your life know how much they mean to you. @MichelleObama, thank you for being a wonderful mother and role model to our daughters and to so many others around the world.

...

(b) Twitter information.

Figure 1: An example of paired WikiData and Twitter information. Relevant text spans with corresponding attribute values are highlighted with the same color.

predicate set, we find WikiData profiles that have Twitter accounts. We further collect Twitter information for each account, including their recent tweets and Twitter metadata, and build models to infer profiles from collected Twitter information, which is solely based on publicly available information and does not involve any additional human annotation efforts.

We first follow Li et al. (2014) to use profile information to generate distant supervised instances and build a sequence labeling-based profile extraction model, similar to Qian et al. (2019). In order to allow open-domain inference, we propose to use attribute names as prompts (Lester et al., 2021) for input sequences to capture the semantics for attribute predicates instead of involving attribute names into the tag set. However, the extraction approach requires that answers must appear in the Twitter context, which ignores some implicit text clues. Therefore, we further propose a prompt-based generation method (Raffel et al., 2020) to infer user profiles, which can additionally produce values that are not straightforwardly mentioned in the Twitter information.

Our statistics show that only a limited number of WikiData attribute values can be directly extracted from Twitter information. Our experiments demonstrate a significant improvement when using the generation-based approach compared to the extraction-based approach, indicating that performing inference instead of pure extraction will be able to obtain more information from tweets. Further analysis shows that the improvement comes mainly from the power of combining extraction and inference on information not explicitly men-

tioned. However, we still find several challenges and limitations for the model to be applied for real-world use, including performances of low-resource attributes, distributional variances between celebrities and normal people, and spurious generation.

Our contributions are summarized as follows:

- To the best of our knowledge, this is the first work to explore open-domain Twitter user profiles.
- We create a new dataset for user profile inference from WikiData, providing with rich and accurate off-the-shelf profile information that can facilitate future social analysis research.
- We propose a prompt-based generation-based method for user profile inference that provides a unified view to infer different attributes.

2 Problem Definition and Dataset

In this section, we first define the open-domain user profile inference and then describe the dataset collection in detail.

2.1 Problem Formulation

The ultimate goal of user profile inference is to infer certain attribute value given the Twitter information of a user. In Twitter, as shown in Figure 1b, we mainly use the collection of recent Twitter tweets from a user u to represent Twitter information, which we denote as

$$\mathbf{X}_{\text{tweet}, u} = [\mathbf{x}_{\text{tweet}, u, 1}, \dots, \mathbf{x}_{\text{tweet}, u, n_{\text{tweet}, u}}],$$

where each $\mathbf{x}_{\text{tweet}, u, i}$ represents a sequence from a single tweet. In addition, we also concatenate the user’s publicly available Twitter metadata (username, display name, bio and location) into a single sequence as complementary user information

Category	#
# predicates	58
# average examples / predicate	12,238
# average candidates / predicate	1,179
# average tokens / answer	1.99
# tweets	13,570,664
# average words per tweet	15.3
# users (train)	106,699
# users (dev)	15,243
# users (test)	30,486

Table 1: Statistics of our collected data from WikiData and Twitter.

$\mathbf{x}_{\text{user},u}$. The final input from Twitter is the combination of user metadata and recent tweets

$$\mathbf{X}_u = [\mathbf{X}_{\text{tweet},u}; [\mathbf{x}_{\text{user},u}]].$$

We then assume that user profiles follow the key-value representation

$$R_u = \{(p_{u,1}, v_{u,1}), \dots, (p_{u,n_r,u}, v_{u,n_r,u})\},$$

where each pair $(p_{u,i}, v_{u,i})$ represents the predicate and value of an attribute. Figure 1a shows an example key-value profile obtained from WikiData.

The model for open-domain user profile inference is to infer the value v of an attribute p from an user u given their Twitter information and a specific attribute predicate with parameter θ

$$f(\mathbf{X}_u, p; \theta) = v.$$

2.2 Dataset Creation

Our dataset consists of WikiData public figure profiles and corresponding Twitter information. An example of paired WikiData profile and Twitter information is shown in Figure 1. We first discuss the collection of WikiData profiles and then discuss the collection of Twitter information.

WikiData processing. WikiData is a structural knowledge base, which can be easily queried with database such as MongoDB² using its dump³. It contains rich encyclopedia information, including information for public figures. Each WikiData entity consists of multiple properties and correspond-

ing claims, which can be considered as the predicate value pairs as shown in Figure 1a⁴.

First, we use WikiData to filter entities that are persons with Twitter accounts. This can be done by checking whether each entity contains the property-claim pair “instance of” (P31) “human” (Q5) and then checking whether the entity includes the property “Twitter username” (P2002). Then we extract the account of those filtered persons using the claim (value) of property “Twitter username” (P2002). If there are multiple claims, we use the first only.

Next, for each entity we check all its properties to build the person’s profile. In Figure 1a, as an example, we can see that the property “occupation” is “politician”. For each property and claim, we only consider their text information, and we use English information only. If there are multiple claims for a property, we use the first one. We drop all properties that do not have an English name for either predicate or value, or properties that do not contain any claims.

Since WikiData profiles usually contain many noisy properties that are not suitable (*e.g.*, blood type) for Twitter user profile inference, we clean the data by 1) filtering extremely low-frequency properties; 2) manually selecting some meaningful and discriminative properties and 3) removing sensitive personal information listed in the Twitter Developer Agreement and Policy, such as political affiliation, ethnic group, religion, and sex or gender⁵. Please refer to Appendix B for the complete list of properties that we use.

Twitter processing. We collect publicly available Twitter information for users that we gather from WikiData, as shown in Figure 1b. The Twitter information consists of the user’s at most 100 recent publicly available tweets, as well as their metadata that includes username, display name, bio (a short description that a user can edit in their profile) and location. We remove all web links and hashtags from those tweets.

Statistics. We collect more than 168k public figures from Wikidata and filter out users whose Twitter accounts are no longer accessible. We obtain about 152K users with 13 million tweets in total. We randomly split the users into train, development

⁴Please refer to <https://www.mediawiki.org/wiki/Wikibase/DataModel> for further details of Wikibase DataModel.

⁵<https://developer.twitter.com/en/developer-terms/agreement-and-policy>

²<https://www.mongodb.com/>

³<https://dumps.wikimedia.org/wikidata/wiki/entities/>

Category	Our Data	Li et al. (2014)	Fang et al. (2015)
# predicates	58	3	6
# users	152K	10.6K	2.5K
# values	709K	10.6K	15K
# tweets	13M	39M	846K

Table 2: Comparison between datasets. Our data contains a diverse set of attributes, with more users and values obtained from WikiData.

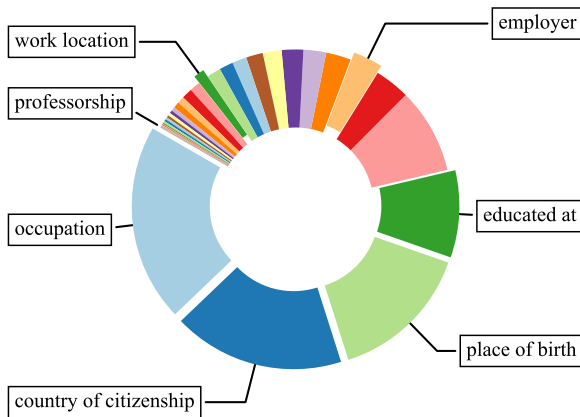


Figure 2: The long tail distribution of different predicates. A few predicates have many examples while most other predicates only have limited examples.

and test sets by 7:1:2. The detailed statistics are shown in Table 1. We compare it with previous work such as Li et al. (2014) and Fang et al. (2015), demonstrated in Table 2. We find that our dataset contains much more diverse predicates compared to Li et al. (2014) and Fang et al. (2015). We also have a much larger number of users and attribute values compared to the previous work. Although Li et al. (2014) contains more tweets than ours, they only consider the extraction setting, and most of the tweets in their datasets are negative samples.

Long tail distribution of predicates. As shown in Figure 2, the number of examples per predicate follows a long tail distribution. Only a few predicates have many training examples, while most appear only partially in the user’s entity list. This raises a huge challenge for us to develop a good model to utilize and transfer the knowledge from rich-resource predicates to low-resource predicates. We discuss the details in the following section.

...	On	behalf	of	the	United
○	○	○	○	○	B
Nations	,	Secretary	-	General	...
I	○	○	○	○	○

Figure 3: An example tweet and tag sequence for attribute `employer` and value `United Nations`.

3 Methods

In this section, we discuss our methods for open-domain Twitter user profile inference. First, we introduce an extraction-based method that largely follows the principle from Li et al. (2014) and Qian et al. (2019). Then we discuss our proposed prompt-based generation approach that provides a unified view to infer different attribute values, and can further infer values that do not appear in the Twitter context.

3.1 Extraction-based Method

We follow Li et al. (2014) and Qian et al. (2019) to generate distantly supervised training instances for user profile extraction. Since our problem is open domain, we propose using attribute predicates as prompts in input sequences and perform sequence labeling over them. This method can be divided into three steps: label generation, modeling, and result aggregation.

Label generation. Distant supervised labeling assumes that if a user u ’s profile contains attribute value v , we can find mentions in their Twitter information expressing the value.

Specifically, we consider each sequence x_i in X_u independently. For each attribute predicate-value pair (p_j, v_j) in u ’s profile, we construct a tag sequence t_{i,p_j} for x_i and the predicate p_j . For a span $[x_b, \dots, x_e]$ that matches v_j , we make

$$t_{i,p_j,b} = B,$$

$$t_{i,p_j,b+1} = \dots = t_{i,p_j,e} = I.$$

If a position k does not match the value, then $t_{i,p_j,k} = O$. For simplicity, we use exact string matching between v_j and spans in the sequence. An example tag sequence is shown in Figure 3.

Modeling. Sequence labeling tasks usually include the label name in the tag set (e.g. B-PER for the beginning of a mention representing a person; Lample et al., 2016). In the open-domain profile inference setting, we have numerous attributes and

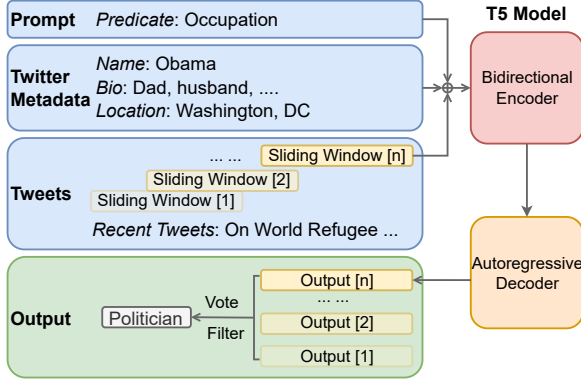


Figure 4: The workflow of the generation-based method, which takes the combination of predicate, Twitter metadata and a window of tweets as input for a T5-based model, and aggregate the window-level results into user level using majority vote.

many of them have only a few instances as shown in Figure 2, which are not sufficient to be considered as separate tag labels.

Therefore, we propose to use prompt-guided sequence labeling, where we append the attribute predicate p to the front of the sequence as the prompt as follows:

$$[\text{CLS}] p [\text{SEP}] \mathbf{x}_i$$

Then we perform sequence labeling on the second part of the input \mathbf{x}_i using the generated labels. We use RoBERTa (Liu et al., 2019) as the backbone encoder, and we denote the last hidden states of \mathbf{x}_i by $\mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_n]$ where n represents the length of \mathbf{x}_i . The probability of predicted labels is

$$P(t_{i,p,k} | \mathbf{x}_i, p) = \text{softmax}(\mathbf{W}_h \mathbf{h}_k + \mathbf{b}_h) \in \mathbb{R}^3,$$

where k represent the position in \mathbf{x}_i .

During training, we randomly drop negative instances that do not contain any B labels to keep the positive-negative sample ratio steady.

Result aggregation. During inference, for each user, we first perform sequence labeling on every sequence predicate pair exhaustively. Then we aggregate sequence-level labeling results into user-level results. For each attribute predicate, we select the span that has the largest averaged logit as the final answer.

3.2 Generation-based Method

Extraction-based methods suffer from the fact that attribute values must appear in the Twitter context.

Instead with user profile inference, it is very likely that we cannot directly find those values in the context and therefore need to infer them using implicit evidence. To address this issue, we propose to use the conditional generation method, which has been shown to be effective in both extracting input information (Raffel et al., 2020; Li et al., 2021) and performing inference and summarization (See et al., 2017; Alshomary et al., 2020). The overall framework is illustrated in Figure 4.

Modeling. We use T5 (Raffel et al., 2020), a generative transformer based model, to directly generate the answer given the predicate. Similar to the extraction-based method, to address the long-tail distribution problem we use the attribute predicate as prompt at the beginning of the input sequence, which can capture rich semantics of those open-domain attribute predicates, especially when the attribute predicate lacks examples in the data. Specifically, the input is the concatenation of prefix predicate (e.g. `predicate:occupation`), user’s Twitter metadata, and the sequence of tweets that the user has recently published. We train the model to generate the attribute value (y_1, \dots, y_n) by minimizing the cross-entropy loss:

$$\mathcal{L}_{CE} = -\frac{1}{n} \sum_{i=1}^n \log p(y_i | \mathbf{y}_{<i}, \mathbf{x}),$$

where \mathbf{x} is the input to the model and n represents the length of the output sequence.

Since we have at most 100 recent tweets of each user whose total length normally exceeds the limit of the model, we use sliding window and divide recent tweets organized in chronological order into different windows where each window can represent information within a time range. Then we train the model on these divided examples separately. Each example contains the same prefix predicate and Twitter metadata but uses different parts of the tweets to infer the attribute value.

Result aggregation. During inference, we use the same sliding window strategy and divide the input into different examples to make predictions independently. Then, similar to the extraction-based method, we aggregate those window-level predictions into a user-level prediction. We count the occurrences of each predicted text for a predicate and then use majority vote to find the aggregated result of that predicate.

Model	Development Set			Test Set		
	Precision	Recall	F ₁	Precision	Recall	F ₁
Random	0.22	0.22	0.22	0.23	0.23	0.23
Majority	14.56	14.56	14.56	14.19	14.19	14.19
Extraction	18.36	9.69	12.69	18.39	9.80	12.79
Generation	59.05	43.71	50.23	58.73	43.40	49.92

Table 3: System performance (%) on our constructed open-domain Twitter user profile inference dataset.

Result filtering. The generation-based method aggressively generates output without estimating whether the generated output is spurious. Therefore, it is important to filter those incorrect predictions during inference.

After result aggregation, we first take the product of probability for each generated token as the score for each aggregated prediction, and then use the averaged score over all aggregated predictions as the confidence score for the aggregated result. A low confidence score indicates that the model cannot determine whether the prediction is valid.

For each predicate, we search the best threshold and set predictions with confidence scores lower than threshold as “no prediction”. We consider all predicted confidence scores from the development set as candidate thresholds and choose the threshold that yields the best performance on the development set. The best searched threshold is then directly applied to filter results on the test set.

4 Experiments

In this section, we conduct experiments on our constructed dataset and user profile extraction dataset (Li et al., 2014). Then we provide a qualitative analysis and discuss the remaining challenges.

4.1 Experimental Setup

We use `roberta-base`⁶ as the base model for the extraction-based model, as it demonstrates its effectiveness on multiple sequence labeling tasks. We use `t5-small`⁷ for the generation-based model, which has much fewer parameters than `roberta-base`. Please refer to Appendix A for a detailed hyperparameter setup and estimated training and inference time.

Evaluation metric. We choose user-level F₁ as our evaluation metric. Specifically, we suppose

⁶<https://huggingface.co/roberta-base>

⁷<https://huggingface.co/t5-small>

Model	Precision	Recall	F ₁
Random	0.26	0.26	0.26
Majority	4.77	4.77	4.77
Extraction	72.14	71.47	71.80
Generation	77.64	68.60	72.84

Table 4: System performance (%) on the subset of the test set that we can find occurrences of attribute values in Twitter context.

a user profile consists n different attributes. We use $C(\cdot)$ to represent the count of different types of output. $C(\text{no prediction})$ refers to the count of “no predictions” and $C(\text{correct prediction})$ refers to the count of predictions that match the WikiData profile. Then we obtain the user-level F₁ as follow:

$$\begin{aligned} \text{precision} &= \frac{C(\text{correct prediction})}{n - C(\text{no prediction})} \\ \text{recall} &= \frac{C(\text{correct prediction})}{n} \\ F_1 &= 2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \end{aligned}$$

We consider the prediction go be valid when it identically matches the ground truth. We do not use entity-level or tag-level F₁ as Qian et al. (2019) because it is not applicable to the generation model. We do not use generation-based metrics (*e.g.*, BLEU) because we observe that most predictions are very short. In addition, compared to no prediction, we want to penalize wrong predictions more. In F₁, the basis of precision does not include “no prediction” results from models while it still has a penalty for wrong predictions.

4.2 Results

4.2.1 Results on User Profile Inference

The main results are shown in Table 3. The random result means that predictions are uniformly

Category	EDUCATION			JOB		
	Precision	Recall	F ₁	Precision	Recall	F ₁
GraphIE	92.87	79.74	85.77	76.03	61.01	67.66
Generation	94.28	91.40	92.82	78.97	65.78	71.76

Table 5: Results on Li et al. (2014) following the preprocessing as Qian et al. (2019). We re-evaluate the results based on user-level F₁. $p < 0.01$ for both F₁ comparisons.

randomly selected and the majority means that predictions are selected with the values that occur most frequently in the training set. We find that both simple methods perform poorly. Overall, we find that our generation-based method significantly outperforms other methods by a large margin. We also find that the extraction-based method cannot even outperform the majority baseline. The reason is that the majority vote can achieve relatively high accuracy on attributes that have a relatively small number of candidates, or one specific candidate takes a large portion of the data, while we cannot find corresponding occurrences of some of those attributes in the Twitter context.

To verify the above claim, we perform another test on a subset of the test set data, for which we can find corresponding occurrences of attribute values in the Twitter context. We find that only 13.56% of the test data can find those value occurrences, which indicates that the majority of the data cannot be directly extracted from Twitter context. The results are shown in Table 4. By comparing the results with overall results, we can find that both extraction and generation systems can get better performance on the subset that we can find occurrences of attribute values. We find that the extraction method performs quite closely to the generation-based method in this setting, though the generation-based method performs better on precision and F₁ and the extraction-based method better on recall. This result indicates that when attribute values occur in Twitter context, the extraction model can effectively extract them, while the generation-based method can additionally infer values that are not included in the Twitter content.

4.2.2 Results on User Profile Extraction

We conduct additional experiments on the profile extraction dataset from Li et al. (2014), where we can provide a direct comparison between our generation-based model and previous work. We follow the same preprocessing as Qian et al. (2019)

Model	Precision	Recall	F ₁
Our model	59.05	43.71	50.23
-threshold	45.95	45.95	45.95
-aggregation	57.39	43.35	49.39
-metadata	53.59	40.45	46.10

Table 6: Effects (%) of result filtering (-threshold), result aggregation (-aggregation) and Twitter metadata on development set. $p < 0.01$ for F₁ comparisons.

on EDUCATION and JOB. We make two changes to our generation-based model for this dataset. 1) This dataset does not contain a timestamp for each tweet, so we use each tweet as an independent sample instead of the sliding window strategy. 2) This dataset is designed for extraction, so for tweets from which the answers cannot be extracted we train the generation model to output “no prediction”.

The experiment results are shown in Table 5. We compare with GraphIE (Qian et al., 2019), one of the state-of-the-art model on this dataset. We reproduce the results from their script⁸ and re-evaluate on user-level with majority vote. We use the averaged results over 5-fold cross validation as Qian et al. (2019). The results show that our model can significantly outperform GraphIE on both EDUCATION and JOB attributes, which indicates that even if the attributes are limited, the generation-based method can still achieve promising performance.

4.3 Ablation study

We conduct an ablation study on two of our components, result filtering and result aggregation, on our profile inference data, as shown in Table 6. We find that result filtering can successfully filter spurious results by improving over 13% on precision, while only dropping about 2% on recall. We also find that result aggregation improves both precision and recall, indicating that we can obtain better

⁸<https://github.com/thomas0809/GraphIE>

... One of the proudest moments of my career being the flag bearer at the Olympics for my home country of Denmark! ...		
Attribute	Value	
country of citizenship	Denmark	✓
... It is going to be February 9, 2022 in Royal Arena against my great friend! ... Beach bod/Mom bod Mommy daughter pool time 2 months with our little angel she clearly enjoyed her first tennis lesson ...		
Attribute	Value	
occupation	tennis player	✓
... bio: Member of the European Parliament ... Still unclear about strategic autonomy. We can't flip a coin when deciding about 2% GDP for Need a clear mechanism for EU intervention. My view in on needs to get a chance to win 5G race...		
Attribute	Value	
occupation	politician	✓
... Can't wait this match vs Brock! Wow...Amazing match Undertaker def 21-0 ... Thanks for a great show. And I CAN WRESTLE ...		
Attribute	Value	
occupation	professional wrestler	actor

Figure 5: Example window-level predictions from generation-based model with their context.

inference by using a larger Twitter context. Twitter metadata also provides rich information about the user’s background. We train and evaluate another model without Twitter metadata, and find that we see a significant performance drop. But we still find that many attributes inferred by the model are not dependent on those metadata.

4.4 Qualitative Analysis

Figure 5 demonstrates four window-level predictions from generation-based model with relevant input context. The first case shows that the model can directly copy relevant information from context. The second and third cases show that the model can infer the information based on the context. The last case shows an error that the model does not fully utilize the information provided by “wrestle” and generates incorrect information, possibly affected by the other word “show”. This case indicates the importance of background information for a specific attribute value.

4.5 Remaining Challenges

Although achieving improvement on open-domain attribute inference, we still find that the model’s performance on attributes with low training samples is generally much lower than on attributes with rich samples. It is still under investigation for better generalization on these low-resource attributes.

WikiData provides rich profiles for many Twitter users. However, the distribution of these Twitter

users with WikiData profiles may not align with the need for downstream tasks. For example, most people with WikiData profiles are celebrities, such as politicians and athletes, which lacks information for general occupations such as farm worker.

The granularity of prediction results is also another important directions to investigate. We observe some cases that the prediction and the groundtruth are in different levels of granularity. For example, the groundtruth can be “Tokyo” while the prediction may be “Japan”. Therefore, it is also important to address this issue with both better modeling as well as evaluation.

We consider that the model can predict all collected attribute values because we have manually selected meaningful and discriminative properties from WikiData during dataset construction. However, it is still possible that a specific property value cannot be detected well based on Twitter content, leading to spurious generation output. For example, if a user is a medical doctor but did not discuss any medical information on Twitter, the occupation is very hard to predict. It is still important to further investigate this “cannot predict” cases in both dataset construction and model design.

5 Related Work

User Profile Inference. One line of user modeling research focuses on profile inference or extraction. Previous work on user profile inference focuses on some specific attributes such as gender (Rao et al., 2011; Liu et al., 2012; Liu and Ruths, 2013; Sakaki et al., 2014), age (Rosenthal and McKeown, 2011; Sap et al., 2014; Chen et al., 2015; Fang et al., 2015; Kim et al., 2017), and political polarity (Rao et al., 2010; Al Zamal et al., 2012; Demszky et al., 2019). They often consider them as multi-class classification problems. Most of these methods use the context of those social media posts. Alternatively, user name and profile in social media (Liu et al., 2012; Liu and Ruths, 2013), part-of-speech and dependency features (Rosenthal and McKeown, 2011), users’ social circles (Chen et al., 2015) and photos (Fang et al., 2015) have been explored as additional important features for different attribute inference. But those classification settings have a pre-defined ontology or label set, which is difficult to extend to other attributes.

In addition to classification-based methods, there are also graph-based (Qian et al., 2017), distant supervision-based and unsupervised extrac-

tion (Huang et al., 2016). Compared to the classification method, extraction-based methods are capable of identifying attributes with a large ontology. But they rely on entities from the context as candidates, which limits the scope of the attributes that occur frequently in the social media context.

Our open-domain Twitter user profile inference uses a larger predicate set and data than previous work. We further propose the generation-based approach, which addresses the limited scope.

Another line of user modeling research focuses on leveraging behavior signals (Kobsa, 2001; Abel et al., 2013) or building implicit user representations (Islam and Goldwasser, 2021, 2022), which is more distantly related to our problem.

Sociolinguistic variation. The intuition of inferring user attributes from their posts aligns with sociolinguistic variation in which people investigate whether a linguistic variation can be attributed to different social variables (Labov, 1963). Computational efforts to discover these relationships include demographic dialectal variation (Blodgett et al., 2016), geographical variation (Eisenstein et al., 2010; Nguyen and Eisenstein, 2017), syntactic or stylistic variation over age and gender (Johannsen et al., 2015), socio-economic status (Flekova et al., 2016; Basile et al., 2019).

6 Conclusion

In this paper, we first explore open-domain Twitter user profile inference. We use the combination of WikiData and Twitter information to create a large-scale dataset. We propose to use a generation-based method with attributes as prompts and compare it with the extraction-based method. The result shows that the generation-based method can significantly outperform the extraction-based method on open-domain profile inference, with the ability to perform both direct extraction and indirect inference. Our further analysis still finds some of the errors and remaining challenges of the generation-based method, such as degraded performances for low-resource attributes and spurious generation, which reveals the limits of our current generation-based user profile inference model.

Limitations

Besides the technical challenges discussed in Section 4.4-4.5, limitations of this work also include the issue of data imbalances that some attributes

may have imbalance distributions. For example, we may find significantly more profiles with the country of citizenship as United States than any other countries, which may have a negative impact on generalization, especially when the distributions of training and inference diverge. Similarly, the distributional variances discussed in Section 4.5 indicate that the prediction results for non-celebrity distributions should be carefully adjudicated. The degraded performances on low-resource attributes also indicate that the prediction results may be unreliable when performing inference on attributes without enough training data.

In this paper, we assume that the attributes are already given. However, many WikiData attributes are not applicable to everyone. For example, attributes such as “position played on team” may be specific to athletes. Therefore, it is also important to investigate how to automatically detect applicable attributes for certain users.

In this work, we use at most 100 recent tweets and aggressively create training and inference examples between each attribute and those tweets. Since we use sliding window on the collected tweets, involving more tweets in training or inference may significantly increase the time cost.

Ethics Statement

The goal of this paper is to extend Twitter user profile inference from limited attributes to the open domain. We hope that this work will help to illustrate how people express their attributes both explicitly but especially also implicitly through their social media posts. We also believe that the NLP community has to produce detailed information about the potential, pitfalls, and basic limitations of profile inference methods so that we can establish standards to facilitate proper use of these technologies, as well as be vigilant and effective at combating nefarious applications.

Data and model biases. To mitigate potential distributional biases, we exhaustively collect entities from WikiData without selecting certain groups of users. However, we acknowledge that the collective information may still contain unintentional social biases. As an example, one of the potential issues is that people who have WikiData profiles are public figures, which may not reflect the actual distribution over general populations (*e.g.*, occupation). Besides, as in Abid et al. (2021), large language models themselves may contain biases.

WikiData is constantly edited by a large number of WikiData contributors and maintainers. Although we try to make our study as representative as possible, it is possible that a statement from WikiData may not reflect the preception from certain groups or individual (Shenoy et al., 2022). We would like stakeholders to be aware of these issues and we urge stakeholders to first investigate the effect of potential issues before drawing any conclusions for any individual or social group using this work.

Proper use v.s. improper use. The major difference between proper use and improper use is whether the use case follows necessary legal and ethical regulations or framework. For example, Williams et al. (2017) propose an ethical framework based on users’ consent to conduct Twitter social research. If the information is not publicly available, one must obtain consent. Opt-out consent can be used when the information is not sensitive, otherwise opt-in consent is required. With proper regulations, this work can be used to enhance personalized user experience, investigate what stakeholders to know to effectively protect personal information.

Sensitivity of personal information. In this work we follow Twitter Developer Agreement and Policy and remove sensitive personal information. But it is still possible to infer sensitive information indirectly. For example, “candidacy in election” may be possibly used to infer political affiliation although the affiliations are generally public for those people. Similarly, personal pronouns, widely present in tweets, may also be used to infer gender. Furthermore, combinations of various sources might allow personal identification (Sweeney, 2000a,b). Even though we do not use private information in our work, based on our results, we speculate that there are unobserved risks of privacy loss for using Twitter. Therefore, We ask that future work should fully comply with regulations and any non-public or private results should be properly protected (Kreuter et al., 2022).

We have set up the following protocol to ensure the proper use and to prevent adverse impact:

- We believe that increasing the transparency of the pipeline can help prevent potential social harm. We plan to release all necessary resources for research reproduction purposes so that others can audit and verify it and prevent overestimation of the model. We also provide a complete list of attributes in Table 7 to increase the transparency.

We are open to all further explorations that can prevent unintended impacts.

- Our constructed dataset for profile inference research is drawn solely from publicly available WikiData and Twitter, where the ethical consideration should be similar to other work using encyclopedia resources such as (Sun and Peng, 2021). Furthermore, according to WikiData: Oversight, non-public personal information are monitored and removed by Wikidata. According to Wiki-Data Term of Use, we can freely reuse and build upon on WikiData. According to the Twitter Developer Agreement and Policy, we will only release IDs instead of actual content for non-commercial research purposes from academic institutions.
- To ensure the proper use of this work, we will not release the data via a publicly available access point. Instead, we will release the data based on individual request and we will ask for consent that 1) requesters are from research institutions 2) they will follow all the regulations when using our work 3) they will not use the model to infer non-public users unless obtained proper consent from those users.

References

- Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. 2013. Twitter-based user modeling for news recommendations. In *Twenty-Third International Joint Conference on Artificial Intelligence*. Citeseer.
- Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Large language models associate muslims with violence. *Nature Machine Intelligence*, 3(6):461–463.
- Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. 2012. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 6, pages 387–390.
- Milad Alshomary, Shahbaz Syed, Martin Potthast, and Henning Wachsmuth. 2020. [Target inference in argument conclusion generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4345, Online. Association for Computational Linguistics.
- Reinald Kim Amplayo. 2019. [Rethinking attribute representation and injection for sentiment classification](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5602–5613, Hong Kong, China. Association for Computational Linguistics.

- Ravi Arunachalam and Sandipan Sarkar. 2013. [The new eye of government: Citizen sentiment analysis in social media](#). In *Proceedings of the IJCNLP 2013 Workshop on Natural Language Processing for Social Media (SocialNLP)*, pages 23–28, Nagoya, Japan. Asian Federation of Natural Language Processing.
- Krisztian Balog, Filip Radlinski, and Shushan Arakelyan. 2019. [Transparent, scrutable and explainable user models for personalized recommendation](#). In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19*, page 265–274, New York, NY, USA. Association for Computing Machinery.
- David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.
- Angelo Basile, Albert Gatt, and Malvina Nissim. 2019. [You write like you eat: Stylistic variation as a predictor of social stratification](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2583–2593, Florence, Italy. Association for Computational Linguistics.
- Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. [Demographic dialectal variation in social media: A case study of African-American English](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.
- Xin Chen, Yu Wang, Eugene Agichtein, and Fusheng Wang. 2015. A comparative study of demographic attribute inference in twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, pages 590–593.
- Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. 2019. [Analyzing polarization in social media: Method and application to tweets on 21 mass shootings](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2970–3005, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. [A latent variable model for geographic lexical variation](#). In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, Cambridge, MA. Association for Computational Linguistics.
- Quan Fang, Jitao Sang, Changsheng Xu, and M. Shamim Hossain. 2015. [Relational user attribute inference in social media](#). *IEEE Transactions on Multimedia*, 17(7):1031–1044.
- Lucie Flekova, Daniel Preoțiuc-Pietro, and Lyle Ungar. 2016. [Exploring stylistic variation with age and income on Twitter](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 313–319, Berlin, Germany. Association for Computational Linguistics.
- Ido Guy. 2015. [The role of user location in personalized search and recommendation](#). In *Proceedings of the 9th ACM Conference on Recommender Systems, RecSys '15*, page 236, New York, NY, USA. Association for Computing Machinery.
- Chao Huang, Dong Wang, Shenglong Zhu, and Daniel Yue Zhang. 2016. [Towards unsupervised home location inference from online social media](#). In *2016 IEEE International Conference on Big Data (Big Data)*, pages 676–685.
- Tunazzina Islam and Dan Goldwasser. 2021. [Analysis of twitter users' lifestyle choices using joint embedding model](#). In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media, ICWSM 2021, held virtually, June 7-10, 2021*, pages 242–253. AAAI Press.
- Tunazzina Islam and Dan Goldwasser. 2022. [Twitter user representation using weakly supervised graph embedding](#). In *Proceedings of the Sixteenth International AAAI Conference on Web and Social Media, ICWSM 2022, Atlanta, Georgia, USA, June 6-9, 2022*, pages 358–369. AAAI Press.
- Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. [Cross-lingual syntactic variation over age and gender](#). In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 103–112, Beijing, China. Association for Computational Linguistics.
- Sunghwan Mac Kim, Qionгкаi Xu, Lizhen Qu, Stephen Wan, and Cécile Paris. 2017. [Demographic inference on Twitter using recursive neural networks](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 471–477, Vancouver, Canada. Association for Computational Linguistics.
- Alfred Kobsa. 2001. Generic user modeling systems. *User modeling and user-adapted interaction*, 11(1):49–63.
- Anne Kreuter, Kai Sassenberg, and Roman Klinger. 2022. [Items from psychometric tests as training data for personality profiling models of Twitter users](#). In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 315–323, Dublin, Ireland. Association for Computational Linguistics.
- William Labov. 1963. The social motivation of a sound change. *Word*, 19(3):273–309.

- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. [Neural architectures for named entity recognition](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.
- Brian Larson. 2017. [Gender as a variable in natural-language processing: Ethical considerations](#). In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. [The power of scale for parameter-efficient prompt tuning](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jiwei Li, Alan Ritter, and Eduard Hovy. 2014. [Weakly supervised user profile extraction from Twitter](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 165–174, Baltimore, Maryland. Association for Computational Linguistics.
- Sha Li, Heng Ji, and Jiawei Han. 2021. [Document-level event argument extraction by conditional generation](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908, Online. Association for Computational Linguistics.
- Wendy Liu and Derek Ruths. 2013. What’s in a name? using first names as features for gender inference in twitter. In *2013 AAAI Spring Symposium Series*.
- Wendy Liu, Faiyaz Zamal, and Derek Ruths. 2012. Using social media to infer gender composition of commuter populations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 6, pages 26–29.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Kai Lu, Yi Zhang, Lanbo Zhang, and Shuxin Wang. 2015. [Exploiting user and business attributes for personalized business recommendation](#). In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’15*, page 891–894, New York, NY, USA. Association for Computing Machinery.
- Dong Nguyen and Jacob Eisenstein. 2017. [A kernel independence test for geographical language variation](#). *Computational Linguistics*, 43(3):567–592.
- Yujie Qian, Enrico Santus, Zhijing Jin, Jiang Guo, and Regina Barzilay. 2019. [GraphIE: A graph-based framework for information extraction](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 751–761, Minneapolis, Minnesota. Association for Computational Linguistics.
- Yujie Qian, Jie Tang, Zhilin Yang, Binxuan Huang, Wei Wei, and Kathleen M Carley. 2017. A probabilistic framework for location inference from social media. *arXiv preprint arXiv:1702.07281*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Delip Rao, Michael Paul, Clay Fink, David Yarowsky, Timothy Oates, and Glen Coppersmith. 2011. Hierarchical bayesian models for latent attribute detection in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 598–601.
- Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. [Classifying latent user attributes in twitter](#). In *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents, SMUC ’10*, page 37–44, New York, NY, USA. Association for Computing Machinery.
- Sara Rosenthal and Kathleen McKeown. 2011. [Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 763–772, Portland, Oregon, USA. Association for Computational Linguistics.
- Shigeyuki Sakaki, Yasuhide Miura, Xiaojun Ma, Keigo Hattori, and Tomoko Ohkuma. 2014. [Twitter user gender inference using combined analysis of text and image processing](#). In *Proceedings of the Third Workshop on Vision and Language*, pages 54–61, Dublin, Ireland. Dublin City University and the Association for Computational Linguistics.
- Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and Hansen Andrew Schwartz. 2014. [Developing age and gender predictive lexica over social media](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1151, Doha, Qatar. Association for Computational Linguistics.
- Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. [Get to the point: Summarization with pointer-](#)

- generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.
- Xuehua Shen, Bin Tan, and ChengXiang Zhai. 2005. [Implicit user modeling for personalized search](#). In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management, CIKM '05*, page 824–831, New York, NY, USA. Association for Computing Machinery.
- Kartik Shenoy, Filip Ilievski, Daniel Garijo, Daniel Schwabe, and Pedro Szekely. 2022. A study of the quality of wikidata. *Journal of Web Semantics*, 72:100679.
- Jiao Sun and Nanyun Peng. 2021. [Men are elected, women are married: Events gender bias on Wikipedia](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 350–360, Online. Association for Computational Linguistics.
- Latanya Sweeney. 2000a. Simple demographics often identify people uniquely. *LIDAP-WP4, 2000*.
- Latanya Sweeney. 2000b. Uniqueness of simple demographics in the u.s. population. *LIDAP-WP4, 2000*.
- Duyu Tang, Bing Qin, and Ting Liu. 2015. [Learning semantic representations of users and products for document level sentiment classification](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1014–1023, Beijing, China. Association for Computational Linguistics.
- Jaime Teevan, Meredith Ringel Morris, and Steve Bush. 2009. [Discovering and using groups to improve personalized search](#). In *Proceedings of the Second ACM International Conference on Web Search and Data Mining, WSDM '09*, page 15–24, New York, NY, USA. Association for Computing Machinery.
- Denny Vrandečić and Markus Krötzsch. 2014. [Wikidata: A free collaborative knowledgebase](#). *Commun. ACM*, 57(10):78–85.
- Matthew L Williams, Pete Burnap, and Luke Sloan. 2017. [Towards an ethical framework for publishing twitter data in social research: Taking into account users' views, online context and algorithmic estimation](#). *Sociology*, 51(6):1149–1168. PMID: 29276313.
- Jing Yao, Zhicheng Dou, and Ji-Rong Wen. 2020. [Employing Personal Word Embeddings for Personalized Search](#), page 1359–1368. Association for Computing Machinery, New York, NY, USA.
- Yangbo Zhu, Jamie Callan, and Jaime Carbonell. 2008. [The impact of history length on personalized search](#). In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08*, page 715–716, New York, NY, USA. Association for Computing Machinery.

A Detailed Experiment Setup

We use two Nvidia GeForce RTX 3090 GPUs as our computing infrastructure.

Extraction-based method setup. We finetune the model with 10 epochs using AdamW. The learning rate is 5e-5 using the linear scheduler without warmup. The batch size is 128. The hidden size for classification is 768. The positive-negative sample ratio is 1:5. We use tag-level F_1 as Qian et al. (2019) to select the best results on the development set efficiently for a single run. The training time is about 16 hours and inference on test set is about 5 hours.

Generation-based method setup. We fine-tune the model on all sliding window examples for 5 epochs using AdamW. The learning rate is 1e-4 using linear scheduler with no warmup. The batch size is 96. We use gradient clipping with max norm 3 to increase stability during training. We use sliding windows with size 512 and stride 128. We use greedy search during inference. We use exact match to select the best results on development set efficiently for a single run. The training time is about 40 hours and inference on test set is about 3 hours.

B Attribute Descriptions

We provide the descriptions of each attribute from Wikidata in Table 7 to facilitate the understanding of attributes and mitigate the potential impact from dataset biases.

ID	Attribute	Description
P106	occupation	occupation of a person; see also "field of work" (Property:P101), "position held" (Property:P39)
P27	country of citizenship	the object is a country that recognizes the subject as its citizen
P19	place of birth	most specific known (e.g. city instead of country, or hospital instead of city) birth location of a person, animal or fictional character
P69	educated at	educational institution attended by subject
P1412	languages spoken, written or signed	language(s) that a person or a people speaks, writes or signs, including the native language(s)
P641	sport	sport that the subject participates or participated in or is associated with
P108	employer	person or organization for which the subject works or worked
P39	position held	subject currently or formerly holds the object position or public office
P1303	instrument	musical instrument that a person plays or teaches or used in a music occupation
P54	member of sports team	sports teams or clubs that the subject represents or represented
P166	award received	award or recognition received by a person, organisation or creative work
P413	position played on team / speciality	position or specialism of a player on a team
P551	residence	the place where the person is or has been, resident
P1344	participant in	event in which a person or organization was/is a participant; inverse of P710 or P1923
P103	native language	language or languages a person has learned from early childhood
P937	work location	location where persons or organisations were actively participating in employment, business or other work
P3602	candidacy in election	election where the subject is a candidate
P463	member of	organization, club or musical group to which the subject belongs. Do not use for membership in ethnic or social groups, nor for holding a political position, such as a member of parliament (use P39 for that).
P101	field of work	specialization of a person, organization, or of the work created by such a specialist; see P106 for the occupation
P118	league	league in which team or player plays or has played in
P2094	competition class	official classification by a regulating body under which the subject (events, teams, participants, or equipment) qualifies for inclusion
P512	academic degree	academic degree that the person holds
P2416	sports discipline competed in	discipline an athlete competed in within a sport
P1411	nominated for	award nomination received by a person, organisation or creative work (inspired from "award received" (Property:P166))
P361	part of	object of which the subject is a part (if this subject is already part of object A which is a part of object B, then please only make the subject part of object A). Inverse property of "has part" (P527, see also "has parts of the class" (P2670)).
P6886	writing language	language in which the writer has written their work
P6553	personal pronoun	personal pronoun(s) this person goes by
P241	military branch	branch to which this military unit, award, office, or person belongs, e.g. Royal Navy
P410	military rank	military rank achieved by a person (should usually have a "start time" qualifier), or military rank associated with a position

Continue on the next page

Table 7: Attribute Description

ID	Attribute	Description
P2348	time period	time period (historic period or era, sports season, theatre season, legislative period etc.) in which the subject occurred
P710	participant	person, group of people or organization (object) that actively takes/took part in an event or process (subject). Preferably qualify with "object has role" (P3831). Use P1923 for participants that are teams.
P1576	lifestyle	typical way of life of an individual, group, or culture
P2650	interested in	item of special or vested interest to this person or organisation
P740	location of formation	location where a group or organization was formed
P859	sponsor	organization or individual that sponsors this item
P812	academic major	major someone studied at college/university
P8413	academic appointment	this person has been appointed to a role within the given higher education institution or department; distinct from employment or affiliation
P5096	member of the crew of	person who has been a member of a crew associated with the vessel or spacecraft. For spacecraft, inverse of crew member (P1029), backup or reserve team or crew (P3015)
P803	professorship	professorship position held by this academic person
P66	ancestral home	place of origin for ancestors of subject
P112	founded by	founder or co-founder of this organization, religion or place
P3828	wears	clothing or accessory worn on subject's body
P1321	place of origin (Switzerland)	lieu d'origine/Heimatort/luogo d'origine of a Swiss national. Not be confused with place of birth or place of residence
P495	country of origin	country of origin of this item (creative work, food, phrase, product, etc.)
P276	location	location of the object, structure or event. In the case of an administrative entity as containing item use P131. For statistical entities use P8138. In the case of a geographic entity use P706. Use P7153 for locations associated with the object.
P5389	permanent resident of	country or region where a person has the legal status of permanent resident
P1429	has pet	pet that a person owns
P263	official residence	the residence at which heads of government and other senior figures officially reside
P1268	represents	organization, individual, or concept that an entity represents
P3716	social classification	social class as recognized in traditional or state law
P17	country	sovereign state of this item (not to be used for human beings)
P488	chairperson	presiding member of an organization, group or body
P7779	military unit	smallest military unit that a person is/was in
P1716	brand	commercial brand associated with the item
P6	head of government	head of the executive power of this town, city, municipality, state, country, or other governmental body
P159	headquarters location	city, where an organization's headquarters is or has been situated. Use P276 qualifier for specific building
P8047	country of registry	country where a ship is or has been registered

Table 7: Attribute Description

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
4.5 Limitations and Challenges Ethics Statement
- A2. Did you discuss any potential risks of your work?
Ethics Statement
- A3. Do the abstract and introduction summarize the paper’s main claims?
Abstract 1 Introduction
- A4. Have you used AI writing assistants when working on this paper?
Left blank.

B Did you use or create scientific artifacts?

1 Introduction 2 Problem Definition and Dataset

- B1. Did you cite the creators of artifacts you used?
1 Introduction
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
2.2 Dataset Creation Ethics Statement
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Ethics Statement
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Ethics Statement
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
2.2 Dataset Creation Ethics Statement
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
2.2 Dataset Creation

C Did you run computational experiments?

4 Experiments

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4.1 Experiment Setup Appendix A Detailed Experiment Setup

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

3.2 Generation-based method 4.1 Experiment Setup Appendix A Detailed Experiment Setup

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

Appendix A Detailed Experiment Setup

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

3.2 Generation-based method 4.1 Experiment Setup

D Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

No response.

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

No response.

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

No response.

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

No response.

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

No response.