

# Predicting the Demographics of Twitter Users with Programmatic Weak Supervision

Jonathan Tonglet<sup>1†</sup>, Astrid Jehoul<sup>2†\*</sup>, Manon Reusens<sup>1\*</sup>,  
Michael Reusens<sup>3</sup>, Bart Baesens<sup>1,4</sup>

<sup>1\*</sup>Department of Decision Sciences and Information Management, KU  
Leuven, Naamsestraat, 69, Leuven, 3000, Belgium.

<sup>2</sup>Datashift, Oude Brusselsestraat, 14, Mechelen, 2800, Belgium.

<sup>3</sup>Statistics Flanders, Havenlaan, 88 bus 100, Brussels, 1000, Belgium.

<sup>4</sup>Department of Decision Analytics and Risk, University of  
Southampton, 12 University Road, Highfield, Southampton, SO17  
1BJ, United Kingdom.

\*Corresponding author(s). E-mail(s): [manon.reusens@kuleuven.be](mailto:manon.reusens@kuleuven.be);  
Contributing authors: [jonathan.tonglet@student.kuleuven.be](mailto:jonathan.tonglet@student.kuleuven.be);  
[astrid@datashift.eu](mailto:astrid@datashift.eu); [michael.reusens@vlaanderen.be](mailto:michael.reusens@vlaanderen.be);  
[bart.baesens@kuleuven.be](mailto:bart.baesens@kuleuven.be);

†These authors contributed equally to this work.

\*Work done while at KU Leuven.

## Abstract

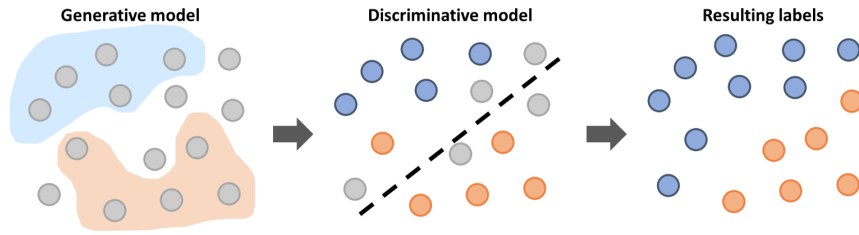
Predicting the demographics of Twitter users has become a problem with a large interest in computational social sciences. However, the limited amount of public datasets with ground truth labels and the tremendous costs of hand-labeling make this task particularly challenging. Recently, programmatic weak supervision has emerged as a new framework to train classifiers on noisy data with minimal human labeling effort. In this paper, demographic prediction is framed for the first time as a programmatic weak supervision problem. A new three-step methodology for gender, age category, and location prediction is provided, which outperforms traditional programmatic weak supervision and is competitive with the state-of-the-art deep learning model. The study is performed in Flanders, a small Dutch-speaking European region, characterized by a limited number of user profiles and tweets. An evaluation conducted on an independent hand-labeled test set shows that the proposed methodology can be generalized to unseen users within the geographic area of interest.

## 1 Introduction

Thanks to its accessible Academic Research API, Twitter is by far the most popular social media data source for computational social science. Various applications of Twitter data include analysing public opinion (Barberá, 2016; Hou et al., 2022), or monitoring health and well-being (Culotta, 2014). Many of those applications are interested in extrapolating their findings to the general population, as an inexpensive and high frequency alternative to traditional surveys (Diaz et al., 2016; Vijayaraghavan et al., 2017). Additionally, researchers are not involved in the data generation process. As a result, Twitter analyses might represent the public opinion more accurately than traditional surveys (Biffignandi et al., 2018). However, using social media data conveys some risks and pitfalls, for example the fact that the demographics of the social media population may differ from those of the general population. Taking the U.S.A. as an example, Mislove et al. (2011) have shown that the Twitter population under-represents women, rural communities, and several ethnic minorities. In Belgium, an analysis conducted by imec showed that, in 2020, 35% of people aged between 16 and 24 used Twitter monthly, against only 9% of the population aged 65 and above (Vandendriessche et al., 2020). As the demographic attributes of Twitter users are not made available through the Academic Research API, demographic prediction has become a critical step of any application aiming to draw conclusions on the general population based on its Twitter counterpart.

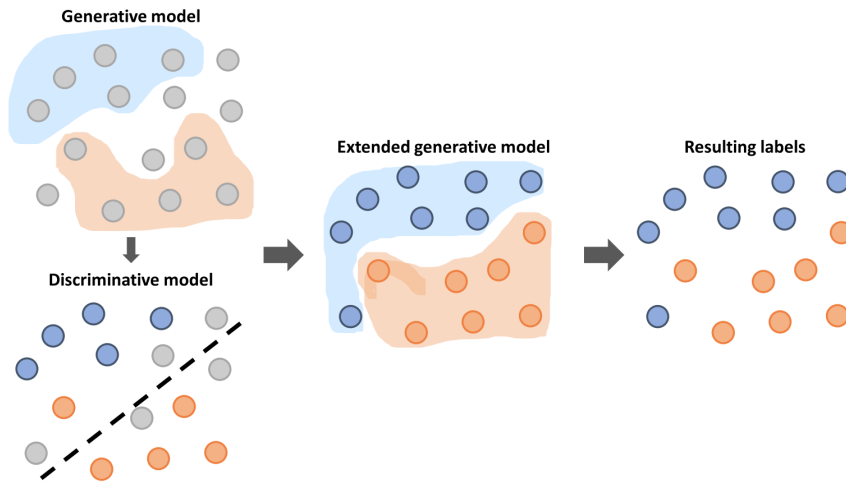
When predicting the demographic attributes of Twitter users, the main challenge is to create an accurate and sufficiently large dataset to train a machine learning (ML) classifier. The number of public datasets with demographic labels is limited and even more so when considering non-English speaking Twitter populations. In addition, acquiring ground truth labels by human labeling is expensive and not scalable. This label scarcity problem, common to several natural language processing tasks, has led to scientific progress in weak supervision. Weak supervision is concerned with relaxing the fully-labeled dataset requirement of traditional supervised learning by collecting a set of weak, i.e. potentially incorrect, labels. Those can be derived from pattern matching, external knowledge sources, expert knowledge, among others. Recently, programmatic weak supervision (PWS) (Ratner et al., 2016; Zhang et al., 2022) has been proposed as a unifying framework of weak supervision, allowing the combination of multiple noisy sources for the same prediction task. To the best of our knowledge, demographic prediction has never been studied within the PWS framework. In traditional PWS, starting from an unlabeled dataset, various weak supervision sources are represented as labeling functions, which potentially overlap and conflict with each other. The labeling functions serve as heuristics to assign weak labels to users. They are optimally combined in a generative model to maximize the coverage of the dataset and the probability that the observed labels occur (Zhang et al., 2022). Afterwards, the weakly labeled data serve as a dataset to train a discriminative model. Once trained, the

discriminative model predicts the final label of both the weakly labeled and unlabeled users, as illustrated on Figure 1.



**Fig. 1** Traditional PWS

In this paper, we propose a new methodology, three-step PWS, as illustrated on Figure 2. Similar to the traditional approach, three-step PWS starts by generating a weakly labeled set from the combination of several labeling functions and trains a discriminative model on it. However, instead of being used as the end model, the discriminative model is added to the set of labeling functions. Then, a second generative model, called the extended generative model, yields the final predictions for all data points. The discriminative model does not leverage features that are already exploited by the labeling functions. Instead, it uses features that are shared by both weakly labeled and unlabeled data points. By doing so, the discriminative model specializes in those data points that could not be covered by the labeling functions of the base generative model, while its predictions on the weakly labeled set are mitigated by the already existing labeling functions, yielding better predictions at the output of the extended generative model.



**Fig. 2** Three-step PWS

The objective of this study is to accurately predict the gender, age category, and location of Twitter profiles without any human labeling effort, using three-step PWS. The scope of our research is limited to a small geographical area called Flanders, the Northern region of Belgium, with a population of 6.5 million. The main language of Flanders is Flemish, a dialect of Dutch. To the best of our knowledge, no demographic prediction studies have been conducted for Flanders. Working with such a small area and population implies that only a limited number of tweets and Twitter accounts are available to train an ML model. This contrasts with previous experiments which have relied on a larger pool of available users to train a demographic prediction model, e.g. Europe (Wang et al., 2019) or the U.S.A. (Barberá, 2016). Hence, a second objective of this study is to identify a sample-efficient method capable of reaching high performance for a small real-world target population.

The proposed method is evaluated on an independent test set with manually appointed ground truth labels. The results achieve promising performance in terms of accuracy and macro F1-score, showing the potential of three-step PWS for demographic prediction on small target populations. Furthermore, the proposed framework is flexible and can easily be extended to new demographic categories and geographical areas of interest, laying the foundation for more experiments. To encourage further research on Twitter demographic prediction, our code is made publicly available<sup>1</sup>.

This paper contributes to the field of demographic prediction as follows. For the first time, demographic prediction is studied as a PWS problem. We provide a new method, three-step PWS, for gender, age category, and location prediction. Three-step PWS outperforms the traditional PWS methodology and is competitive with M3 (Wang et al., 2019), the SOTA deep learning model for gender and age category prediction. We provide the first evaluation and comparison of several labeling functions, ranging from traditional knowledge sources to recent few-shot learners (Radford et al., 2021), and give guidelines on which are most suitable for demographic prediction tasks.

The body of this paper is structured as follows. Section 2 covers related literature on social media demographic prediction, with a particular attention to data labeling strategies. Section 3 outlines the principles of programmatic weak supervision, while section 4 explains the implementation of three-step PWS in the context of demographic prediction. In section 5, experiment results are presented and discussed. The final section contains the conclusion and several suggestions for future research.

## 2 Related work

### 2.1 General aspects

Social media demographic prediction, sometimes referred to as author profiling, is a well-studied problem with more than a decade of scientific research. Most analyses have been conducted on Twitter, due to its user-friendly Academic Research API, but other platforms are considered as well, like Facebook (Rao et al., 2011; Matz et al., 2019). Demographic prediction usually takes the form of a classification task, with

---

<sup>1</sup>All the data collected for our experiments was processed in compliance with the rules of Twitter Academic Search API. To respect users' data privacy, only the code and a small dataset of fictional users are publicly made available on <https://github.com/jtonglet/Demographics-PWS>

coarse bins for continuous target variables like age or income, although regression is also a possibility (Nguyen et al., 2013; Preoțiu-Pietro et al., 2015; Matz et al., 2019). Gender and age are the most frequently studied attributes (Al Zamal et al., 2012; Ikeda et al., 2013; Nguyen et al., 2013; Miranda Filho et al., 2015; Culotta et al., 2016; Vijayaraghavan et al., 2017; Wang et al., 2019; Graells-Garrido et al., 2020). Other attributes can include education level, income, ethnicity, and location. The latter is the only one with a dedicated field in the Twitter user profile. Furthermore, the position of a user can be tracked by an opt-in tweet geolocation service. However, most users do not complete the location field and refuse the geolocation, making location prediction still necessary in practice (Compton et al., 2014; Jurgens et al., 2015; Rahimi et al., 2018). A literature overview is shown in Table 1. For a complete review of the field, we refer to Hinds and Joinson (2018) and HaCohen-Kerner (2022).

Authors, year	Data sources	Input features	Target variables	Classifiers	Performance metrics
Mislove et al. (2011)	Twitter	Text	Gender, Ethnicity, Location	Lexicon-based	F1-score
Pennacchiotti and Popescu (2011)	Twitter	Text, Network	Ethnicity, Party	Tree ensemble	Precision, Recall
Rao et al. (2011)	Facebook	Text	Gender, Ethnicity	BHM	Accuracy
Al Zamal et al. (2012)	Twitter	Text, Network	Age, Gender, Party	SVM	Accuracy
Ikeda et al. (2013)	Twitter	Text, Network	Age, Gender, Location	SVM	F1-score
Nguyen et al. (2013)	Twitter	Text	Occupation, Hobby, Marital status	RLM	Accuracy, F1-score
Compton et al. (2014)	Twitter	Network	Location	Label propagation	Pearson $\rho$ , MAE
Li et al. (2014)	Twitter	Text, Network	Spouse, Education	Global and local distant supervision	Median Error, Mean Error
	Google Plus, Freebase		Occupation		F1-score
Chen et al. (2015)	Twitter	Text, Network	Age, Gender	SVM	Precision, Recall
		Image	Ethnicity		Accuracy, F1-score
Preoțiu-Pietro et al. (2015)	Twitter	Text	Occupation	Gaussian process	Precision, Recall, AUC
Preoțiu-Pietro et al. (2015)	Twitter	Text	Income	Gaussian process	Accuracy
(Miranda Filho et al., 2015)	Twitter	Text	Age, Gender	Naive Bayes, SVM	Pearson $\rho$
			Social Category	Tree Ensemble	F1-score, Accuracy
Barberá (2016)	Twitter, Voting records	Text, Network	Age, Gender, Income	RLM	Accuracy
			Ethnicity, Party, Vote intention		Precision, Recall
Culotta et al. (2016)	Twitter, Quantcast	Text, Network	Age, Gender, Ethnicity, Income	RLM	Correlation, F1-score
			Education, Party, Parental status		
Ardehaly and Culotta (2017a)	Twitter, Census	Text, Image	Income	Neural network	F1-score
Vijayaraghavan et al. (2017)	Twitter	Text, Network, Image	Age, Gender, Location, Party	Neural network	F1-Score
Aletras and Chamberlain (2018)	Twitter	Text, Network	Income, Occupation	RLM, SVM	Accuracy
				Gaussian process	Pearson $\rho$ , MAE
Rahimi et al. (2018)	Twitter	Text, Network	Location	Neural Network	Accuracy, Median Error
					Mean Error
Matz et al. (2019)	Facebook	Text	Income	RLM	Pearson $\rho$
Pan et al. (2019)	Twitter	Text, Network	Occupation	Neural Network	Accuracy
Wang et al. (2019)	Twitter, Wikipedia, IMDB	Text, Image	Age, Gender, is_company	Neural network	F1-score
Graells-Garrido et al. (2020)	Twitter	Text, Network	Age, Gender, Location	Tree ensemble	Precision, Recall
López-Monroy et al. (2020)	Twitter	Text	Gender, Language Variety	SVM	Accuracy
Wood-Doughty et al. (2021)	Twitter	Text	Race, Ethnicity	RLM, Neural network	Accuracy, F1-score
Suman et al. (2021)	Twitter	Text, Image	Gender	Neural network	Accuracy

**Table 1** Review of demographic prediction methods for social media data.

Apart from target variables, existing works differentiate themselves with the type of input features they use. The Academic Research API provides a combination of structured user metrics, tweet metrics, raw text, and image data. While recent advances in deep learning allow to provide raw data directly as input to a neural network, most classifiers require hand-crafted features as a preprocessing step.

The two main sources of text features are tweet content and profile metadata, such as name, screen name, profile description, and profile location. Some researchers, like Wood-Doughty et al. (2021), retrieve hundreds of the user’s past tweets, while others, like Wang et al. (2019), exclusively analyze the profile metadata to avoid expensive tweet queries. Authors have converted raw text into meaningful features using bag of words (BoW) (Barberá, 2016; Culotta et al., 2016), word embeddings (Wang et al., 2019; Wood-Doughty et al., 2021), or topic models (Pennacchiotti and Popescu, 2011; Preoțiu-Pietro et al., 2015,?; Matz et al., 2019). BoW provides a vector representation of a text sequence, where each cell value typically contains the number of occurrences

of a term (words, emojis, hashtags, or URLs) in the sequence (Barberá, 2016; Graells-Garrido et al., 2020). Multiple text sequences can then be concatenated to form a matrix. As most textual terms are seldom used, the resulting matrix suffers from high sparsity. One way to deal with this problem is to only keep the most frequent and important terms, determined by their term frequency-inverse document frequency (TF-IDF) score (Barberá, 2016). An alternative is to represent textual terms as low-dimensional vectors called word embeddings. Word embeddings are denser than BoW vectors and incorporate a notion of word similarity, computed as the cosine similarity between two Embeddings. Finally, topic models identify shared topics in a corpus of text sequences. Topics are then used as predictors of the user’s demographic attributes. One such topic model is Top2Vec (Angelov, 2020), which maps tweets and words to an embedding space, where clusters are indicative of semantic similarity and of the presence of a common topic.

A set of network features can be derived from the different interactions between users, including mention, follow, quote, or retweet relationships. The core concept is to represent the community of Twitter users as a directed graph  $G = (V, E)$ , where vertices  $V$  represent users and directed edges  $E$  represent the relationship between two users. Many researchers make the assumption that the Twitter network is homophilic, implying that users are more likely to share the sociodemographic attributes of their neighbors (McPherson et al., 2001; Al Zamal et al., 2012; Pan et al., 2019). This assumption allows to derive interesting features from the network structure. In Barberá (2016), verified Twitter accounts of celebrities are used as a demographic predictor of their followers. Graells-Garrido et al. (2020) use the adjacency matrices of retweets, mentions, replies, and quotes as feature matrices. Aletras and Chamberlain (2018) and Pan et al. (2019) leverage graph embeddings to obtain dense vector representations of a user’s neighborhood.

Image data has mainly been exploited in raw format with deep learning algorithms, which do not require time-consuming feature engineering (Ardehaly and Culotta, 2017a; Vijayaraghavan et al., 2017; Wang et al., 2019), with some exceptions (Chen et al., 2015). When available, the user’s profile picture is a rich source of information, assuming that the picture actually portrays the user.

Finally, in addition to ready-to-use features like followers and retweet count, authors have proposed various metadata features, such as the retweeting tendency (fraction of posts that are retweets), the tweet frequency, and the fraction of tweets posted per hour (Rao et al., 2010; Pennacchiotti and Popescu, 2011).

Equivalently, authors have relied upon a large variety of classifiers, ranging from regularized linear models (RLM) to more expressive support vector machines (SVM) and tree ensembles. Furthermore, various deep learning architectures have recently been applied to demographic prediction, including long short term memory networks (Wang et al., 2019) and graph convolutional networks (Rahimi et al., 2018; Pan et al., 2019).

## 2.2 Data labeling

Demographic prediction characterizes itself by a label scarcity problem. For all demographic attributes, except location, there is no way to directly access ground truth values. Therefore, authors have been concerned with elaborating methods to derive a

set of weak proxy-labels or, in the best case, retrieve the true label values. Existing labeling methods in the literature can be divided into three families: supervised learning, weak supervision, and semi-supervised learning. Table 2 displays a non-exhaustive list of papers using these techniques.

Method	Subcategory	Papers
Supervised learning	Hand-labeling	Preoțiuc-Pietro et al. (2015); Chen et al. (2015)
	Recruited panel	Matz et al. (2019)
Semi-supervised learning	Label propagation	Compton et al. (2014); Rahimi et al. (2018)
	Co-training	Ardehaly and Culotta (2017a); Wang et al. (2019)
Weak supervision	Heuristics	Wang et al. (2019); Graells-Garrido et al. (2020)
	Distant supervision	Li et al. (2014); Barberá (2016)
	LLP	Ardehaly and Culotta (2017a,b)

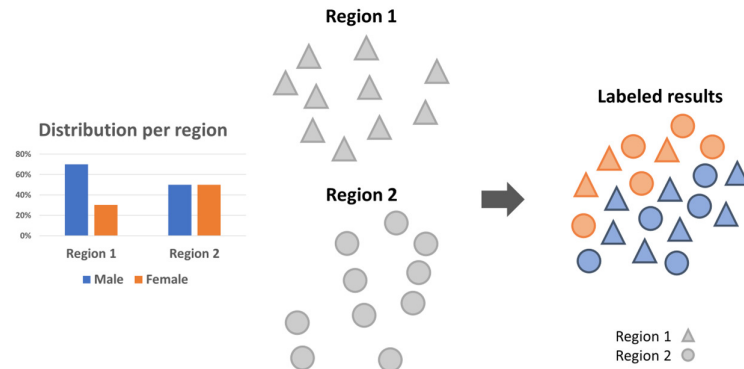
**Table 2** Data labeling methods for demographic prediction.

The first family corresponds to the traditional **supervised learning** setting. An ML classifier is trained on a set of users, usually a small subset of the total user population, for which labels are acquired by hand-labeling or by recruiting a panel. Hand-labeling is typically performed by crowdworkers on Amazon Mechanical Turk (Chen et al., 2015) and FigureEight (Wang et al., 2019). To account for annotator bias, more than one worker is assigned to each unlabeled instance. Confidence in the annotators’ perception can be assessed with inter-annotator agreement metrics, such as Fleiss’  $\kappa$  (Chen et al., 2015) and Krippendorff’s  $\alpha$  (Wang et al., 2019). Unfortunately, hand-labeling becomes more expensive and time-consuming as size of the dataset and required subject matter expertise increase. Therefore, authors tend to only rely on a small quantity of crowd-sourced hand-labels to serve as an evaluation set for their proposed methodology, trained on a larger dataset without ground truth labels (Barberá, 2016; Wang et al., 2019). Another approach is to directly ask Twitter users to join a panel experiment and reveal their true demographic attributes for the purpose of the research (Matz et al., 2019). This is particularly useful when studying complex attributes, such as income or education level. However, this method is the least scalable to large sample of users and it introduces a level of selection bias, given there is no guarantee that the recruited users have an online behavior that is representative for the global Twitter crowd.

**Semi-supervised learning** is an ML framework where the dataset consists of both labeled and unlabeled data. It contains a large range of methods among which two, label propagation and co-training, have been successfully applied to demographic prediction. Label propagation is an iterative algorithm that classifies unlabeled instances based on the labels of their nearest neighbors. Close neighbors are identified with appropriate distance metrics or shortest path algorithms in the latent feature space or a network structure, respectively. It has mainly been used to infer missing location data (Compton et al., 2014; Rahimi et al., 2018). Co-training (Ardehaly and Culotta, 2017a; Wang et al., 2019) is a setting where distinct classifiers are trained on different views of the input data. Unlabeled instances predicted with high confidence by one of the classifiers are given as labeled instances to the others. For example, Wang

et al. (2019) first train a classifier on image data to enlarge the training set of a text classifier trained on profile descriptions.

**Weak supervision** approaches scale to larger datasets by relaxing the requirement of ground truth or human-annotated labels. Instead, they rely on a set of weak, noisy labels generated by heuristics and external knowledge sources. Heuristics assign labels based on self-reported demographic information in user profiles and tweet content (Wood-Doughty et al., 2021). Although heuristics allow for fast labeling, they may introduce significant noise when improperly defined or when self-reported information is inaccurate or fictional. Depending on the demographic attribute, heuristics may, like hand-labeling, require a significant amount of subject matter expertise. Distant supervision accesses external knowledge sources to assign labels. Some authors compare user names to a dictionary of male and female names in the corresponding language (Wang et al., 2019; Graells-Garrido et al., 2020) or reported location(s) with a gazetteer (Graells-Garrido et al., 2020). More elaborated approaches include matching users to their voting registration records that contain their demographic attributes (Barberá, 2016; Grinberg et al., 2019) or retrieving the attributes on other social media accounts belonging to the same person, like LinkedIn (Daas et al., 2016) or Google-Plus (Li et al., 2014). Both heuristic and distant supervision methods have a high risk of selection bias. The last existing weak supervision setting is learning from labeled proportions (LLP), used by Ardehaly and Culotta (2017a,b). In LLP, the unlabeled dataset is split into non-overlapping bags. For each bag, only the percentage of users in each target class is given. In Ardehaly and Culotta (2017a,b), bags correspond to geographic regions for which census data is available. An ML algorithm is trained to label users while preserving the correct proportions of each bag, as illustrated on Figure 3. Their approach makes the assumption that Twitter demographics follow the same distribution as the census, which is unrealistic in practice. Replacing the census by aggregate statistics on Twitter usage per demographic group, if available, would be more appropriate.



**Fig. 3** Learning from Labeled proportions

The boundaries between the aforementioned methods are indubitably blurred. For instance, one could argue that crowd-sourced labels fall into the category of weak supervision sources, when annotators have little expertise and their work is unreliable. Although combining the signal of several weak supervision sources can improve the quality of the weak labels, previous research has focused on only one source per demographic attribute (Barberá, 2016; Culotta et al., 2016; Wang et al., 2019). In the absence of ground truth labels, i.e. in a unsupervised setting, optimally combining multiple weak supervision source requires an appropriate methodology (Ratner et al., 2017). Programmatic weak supervision tackles this problem, as we show in the next section.

### 3 Programmatic weak supervision

Programmatic weak supervision (PWS) was introduced by Ratner et al. (2016) as a unifying framework of weak supervision. PWS aims to recover the unobserved ground truth labels of an unlabeled dataset by combining the efforts of a variety of weak supervision sources. It produces a set of weak labels, partially covering the initial unlabeled dataset, to train downstream ML models on.

In PWS, weak supervision sources are abstracted under the concept of labeling functions (LFs). A LF can be defined as a function  $\lambda : \mathcal{X} \rightarrow \mathcal{Y} \cup \{-1\}$  that maps unlabeled samples  $\mathcal{X}$  to a set of class labels  $\mathcal{Y}$ . Abstaining from labeling an instance is represented by the ‘-1’ label (Zhang et al., 2022). The LFs can be any sort of weak supervision source: keyword and pattern-based heuristics, distant supervision knowledge bases, predictions of weak learners, third-party models, legacy systems, and even labels assigned by crowdworkers (Ratner et al., 2017). By combining multiple LFs, a label matrix

$$L = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \dots & \lambda_{1M} \\ \lambda_{21} & \lambda_{21} & \dots & \lambda_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{N1} & \lambda_{N2} & \dots & \lambda_{NM} \end{bmatrix} = (\lambda_{ij}) \in \mathcal{Y} \cup \{-1\} \quad (1)$$

is obtained, where  $(\lambda_{ij})$  is the label assigned by the weak supervision source  $j \in M$  to sample  $i \in N$ .

The label matrix  $L$  serves as an input to an unsupervised label model, which combines the various LFs to generate a set of weak labels. Some label models yield probabilistic labels as output, i.e. a probability distribution over the possible classes, while others result in one-hot encoded labels. The most basic label model is the majority vote, which simply outputs the mode for each row of the label matrix and abstains in case of ties. However, given LFs are noisy, they often overlap and conflict with each other and may hold hidden dependencies. A simple majority vote may fail to handle those factors. Therefore, a more complex label model, data programming, based on a factor graph, has been proposed in Ratner et al. (2016). Data programming performs a weighted vote of the LFs, where weights are assigned via an unsupervised generative process applied over the LFs (Ratner et al., 2016, 2017). This generative process automatically denoises the values in the label matrix by modeling the LFs interdependencies and by estimating their accuracies. The more dependencies between the LFs, the more the

generative model will outperform the majority vote baseline. The simplest case occurs when there are no dependencies between LFs. Assuming  $m$  LFs  $\lambda_j \in L$ , we define the coverage probability  $\beta_j \in \mathbb{R}^m$  as the probability that LF  $\lambda_j$  labels an object. Similarly, we define the precision probability  $\alpha_j \in \mathbb{R}^m$  as the probability that LF  $\lambda_j$  assigns the correct label to an object. [Ratner et al. \(2016\)](#) represent the distribution of the data programming model for a binary classification task as

$$\mu_{\alpha,\beta}(L) = \frac{1}{2} \prod_{j=1}^m (\beta_j \alpha_j \mathbf{1}_{\{\lambda_j=1\}} + \beta_j(1 - \alpha_j) \mathbf{1}_{\{\lambda_j=0\}} + (1 - \beta_j) \mathbf{1}_{\{\lambda_j=-1\}}) \quad (2)$$

where  $\mathbf{1}_{\{\lambda_j=i\}}$  is the indicator function that takes value 1 when  $\lambda_j = i$  and 0 otherwise. The objective is to find the values of  $\alpha_j$  and  $\beta_j$  that maximize the likelihood function of the generative model, determined with stochastic gradient descent:

$$(\hat{\alpha}, \hat{\beta}) = \arg \max_{\alpha,\beta} \sum_{x \in \mathcal{X}} \log \left( \sum_{y \in \mathcal{Y}} \mu_{\alpha,\beta}(\mathbf{L}(x), y) \right) \quad (3)$$

Each LF  $\lambda_j$  is weighted according to its  $\alpha_j$  and  $\beta_j$ . Afterward, the weighted combination of all LFs yields the predicted weak labels. To remain coherent with previously introduced concepts and their notations, small modifications were made to Equation 2 and 3, compared to their original version in [Ratner et al. \(2016\)](#). For a more detailed explanation of data programming, its formulation in presence of interdependencies, and proofs on its properties, we refer to the seminal paper of [Ratner et al. \(2016\)](#).

Following data programming, other label models have been proposed, such as MeTaL ([Ratner et al., 2019](#)) that tackles the multi-task weak supervision sources, or FlyingSquid ([Fu et al., 2020](#)) which speeds up the weight convergence by changing the optimization method. [Zhang et al. \(2022\)](#) provide a comprehensive literature review of the growing PWS field.

In the following, we use the Python package Snorkel<sup>2</sup>([Ratner et al., 2017](#)) to implement LFs and train generative label models.

## 4 Methodology

The objective of this paper is to predict gender, age category, and location at province level of Flemish Twitter users, using three-step PWS, without any human labeling effort to build the training set. Three-step PWS distinguishes itself from previous demographic prediction weak supervision models in two ways. First, it combines multiple sources, instead of a single one, in a base generative model to generate the weak labels. Second, the discriminative model does not serve as an end-model. Instead, it acts as an additional labeling function with a distinct set of features, specialized in those instances that the base generative model could not label. To make the differences between both processes more clear, we have included Figures 4 and 5 illustrating both pipelines.

For each demographic attribute, we apply the process illustrated on Figure 5, where the three PWS steps are shown in orange. The process starts with a split of the

<sup>2</sup><https://github.com/snorkel-team/snorkel>

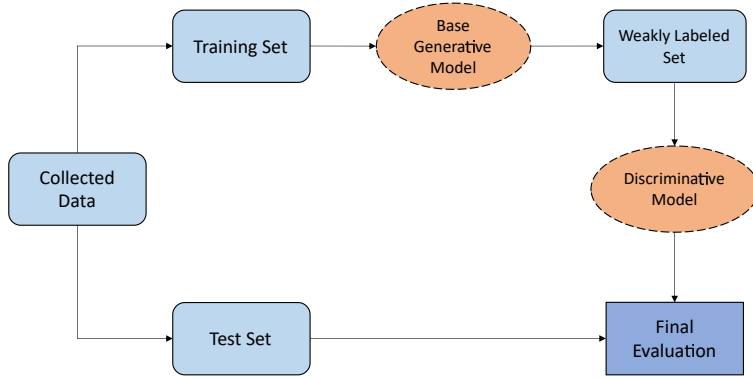


Fig. 4 Traditional PWS process

collected dataset into a training and test set. The latter is set aside and labeled by hand to serve as a final evaluation of the proposed methodology. Then, a first group of LFs are combined in a generative model, referred to as the “base generative model”, to assign weak labels to the training set. Considering these simple LFs cannot fully cover the training set, various discriminative models are trained on the weak labels. The best classifiers are then added as LFs to the generative model to reach a complete coverage of the training set. This second generative model is referred to as the “extended generative model”. A final evaluation of the extended generative model is then conducted on the hand-labeled test set. The following subsections explain the different steps of the methodology in more detail.

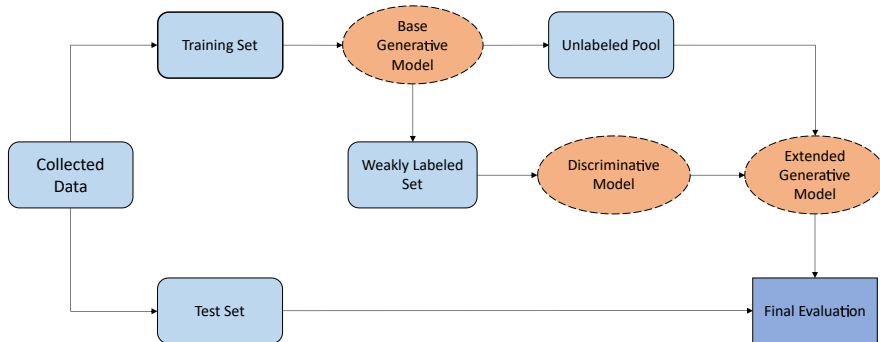


Fig. 5 3-step PWS process for demographic prediction

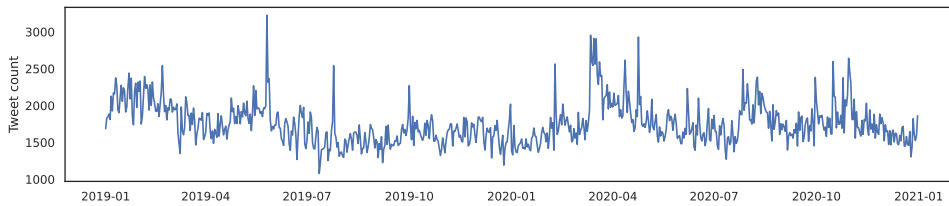
#### 4.1 Target definition

**Gender** prediction is expressed as a classification task with two labels: *Male* and *Female*. Though we recognize that this binary representation is not representative of the gender spectrum, we leave the prediction of a larger range of gender categories

for future work. For **Age**, the same categories as Wang et al. (2019) and Graells-Garrido et al. (2020) are used, namely: *18 and below*, *19-29*, *30-39*, *40 and above*. Seven categories were identified for **Location** prediction. Flanders, our geographic area of interest, is one of the three Belgian regions and its five provinces constitute the first location classes: *Antwerpen (AN)*, *Limburg (LI)*, *West-Vlaanderen (WV)*, *Oost-Vlaanderen (OV)* and *Vlaams-Brabant (VB)*. Another class is created to group users from the two other Belgian regions *Wallonia and Brussels (WA-BX)*. A final class contains *foreign (FO)* users, mainly from the Netherlands.

## 4.2 Data collection

With the Twitter Full-Archive Search<sup>3</sup>, we query all tweets written in Dutch, geolocated in Belgium over the period 2019-2020. No additional query filters are specified. For each tweet, the corresponding user profile data is collected, including the user name and screen name, their profile description, location data, profile image, and public metrics. After removing duplicates, the resulting dataset consists of 28.464 unique user accounts and 1.2 million tweets. This is a fairly small dataset compared to prior works on demographic prediction : 3.98 million, and 1.20 million users for gender and age respectively in Wang et al. (2019), 233132 users for all attributes in Barberá (2016). As shown in Figure 6, the daily tweet volume is quite stable over time, with noticeable peaks in May 2019 for the Belgian Federal elections and in March-April 2020 for the first wave of the Covid-19 pandemic.

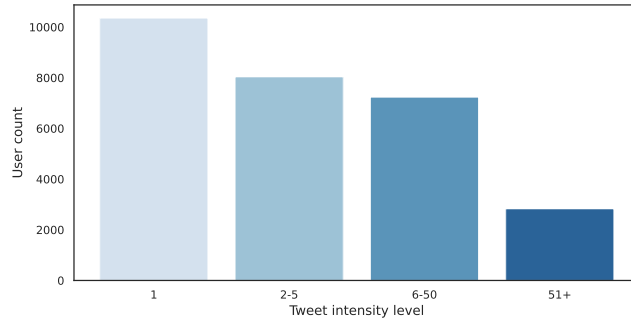


**Fig. 6** Distributions of collected tweets over time

The tweet intensity, i.e. the number of tweets written by a user, is very low. More than a third of the users have only tweeted once and more than half of them have written less than 5 tweets, as shown on Figure 7. Depending on the reliance of the PWS process on the tweet content, this low intensity may impact the performance and deserves to be further investigated. Note that this excludes tweets written by the same users prior to 2019-2020. Retrieving more tweets is a possibility, but requires additional API calls and storage space.

Due to the imprecise geolocation and the shared language, the dataset includes users from Brussels-Wallonia and the Netherlands as well. A handful of users come from more distant countries, mainly people who tweeted once or twice on a trip in Belgium. Hence, the location task includes two additional categories for those users.

<sup>3</sup><https://developer.twitter.com/en/docs/twitter-api/tweets/search/introduction>



**Fig. 7** Users count per tweet intensity level

### 4.3 Hand-labeled Test Set

2% of the user profiles (571 accounts) are randomly set aside to serve as an independent test set. In the absence of ground truth labels, the most reliable proxy is to perform human annotation with an inter-annotator agreement evaluation. Every user profile in the test set is hand-labeled by three independent annotators, based on available account information and under a strict confidentiality agreement. A total number of 14 student annotators are recruited to perform the labeling for this study. When selecting our annotators, trustworthy people were chosen to respect the confidentiality of the data. When a demographic attribute is too ambiguous, annotators can select the “Unknown” option as a way to abstain from labeling. In addition, a special category is introduced to flag company and other organizational accounts. The majority vote among the three annotators is selected as final label. In case of disagreement between all 3 annotators, no label is assigned. In the end, 489 users received a valid (not Unknown) label for gender, 362 for age category, and 413 for location. The inter-annotator agreement, evaluated with Fleiss’  $\kappa$ , is 87% for gender, 57% for age category, and 77% for Location. These moderate to high values confirm that the work of the annotators is reliable. The relatively low value for age category is probably due to the fact that a user’s age is often more hidden on profiles and the perception of age varies among annotators.

### 4.4 Weakly labeled set

The remainder of the collected data constitutes the training set. Table 3 summarizes the collection of LFs used to create the weakly labeled set. As a preprocessing step, accounts in the training set belonging to companies and other organizations are detected with M3<sup>4</sup> (Wang et al., 2019), a pre-trained deep learning model. Setting the threshold at 0.95, M3 predicts that 470 accounts are company profiles. Those accounts are removed from the training set before applying the LFs.

Six heuristic LFs, one for each target value, use discriminative keywords to detect the age and gender, such as “he/him”, “mother” or “retired”, found in profile descriptions. See Appendix B for a list of all keywords used in LFs. In addition, we use regular

---

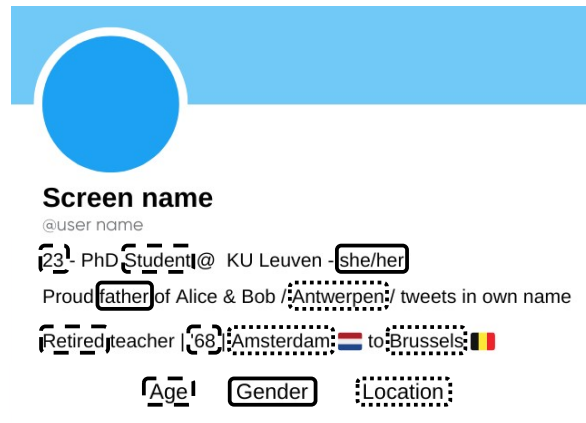
<sup>4</sup><https://github.com/euagendas/m3inference>

Gender	Keywords (heuristics), Names Dictionary (distant supervision), VGG-Face (3 <sup>rd</sup> party model), CLIP 3 <sup>rd</sup> party model)
Age category	
Location	
	Keywords (heuristics) , RegEx patterns (heuristics) Town names (distant supervision), List of ZIP codes (distant supervision)

**Table 3** Type of LFs per task.

expressions (RegEx) to match self-reported age or year of birth in profile descriptions such as “born on 10/12/99”. A total of 28 RegEx LFs are defined. Each heuristic LF is specialized in one attribute value and can either predict that value or abstain.

A dictionary of common male and female first names in Belgium serves as a distant supervision source for gender, while two gazetteers are matched to user profiles for location. One consists of Belgian ZIP codes, the other contains all Belgian and Dutch town names, supplemented by a list of European countries and their capital cities. Both have a mapping to the corresponding location category when there is a match. The town names are applied on both the location profile field and the description field of the user profile, while ZIP codes are applied on the description field only. Figure 8 illustrates how heuristic matching and distant supervision are applied on three fictional profile descriptions.



**Fig. 8** Heuristics and Distant Supervision LFs applied to profile descriptions.

Two third-party models are used as LFs for gender. The first one is a wrapper of the VGG-Face model (Parkhi et al., 2015), provided by the DeepFace<sup>5</sup> library (Serengil and Ozpinar, 2020, 2021). Among other tasks, the model is pre-trained to identify gender from face images. The second model is CLIP<sup>6</sup> (Radford et al., 2021), a zero-shot image classifier that takes a set of text tokens and an image as input and returns the token that is the most associated with the image. For gender prediction, we provide the following text tokens to CLIP: “A man”, “A woman”, and “An object”. When no

<sup>5</sup><https://github.com/serengil/deepface>

<sup>6</sup><https://github.com/openai/CLIP>

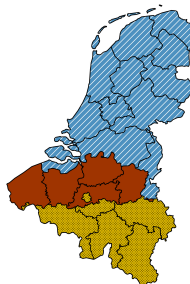
profile image is available or if the profile image URL does not work properly, the two models abstain from labeling.

This is only a selection of the LFs that one can apply to demographic prediction. Less reliable keywords like gender-associated emojis are ignored to favor the precision of the LFs over their recall. Some interesting profile metadata, like the optional user birthday field, are not retrievable from the Academic Research API and therefore not covered by the RegEx LFs. As for the data linkage to other social media accounts, it is avoided for ethical and data privacy reasons.

Next, the base generative model is fitted on the label matrix. Technically, with sufficient precision and full coverage, the generative model could serve as an end model to perform demographic prediction. However, it is highly unlikely that the model reaches full coverage with the current set of LFs.

## 4.5 Discriminative model

Six distinct classifiers are trained for every task separately, namely multiclass and ordinal logistic regression, Random Forest, XGBoost, LightGBM, and Catboost. The location task contains 7 classes and is split further into three subtasks to simplify computations, referred to as Task A, B, and C (illustrated on Figure 9<sup>7</sup>). Task A separates Belgian from Dutch and other non-Belgian accounts, while Task B isolates Flemish users from those of Brussels and Wallonia. Finally, Task C assigns users to one of the five provinces of Flanders. For that reason, the training set of Task C is the Flemish subset of the training set of Task B, which in turn is the Belgian subset of the training set of task A.



**Fig. 9** The three location tasks: Task A and Task B respectively remove users from the blue striped and orange dotted areas. Task C allocates the remaining users to one of the five Flemish provinces in red.

**Logistic regression** (LR) is selected to be used as a benchmark model, given its computational efficiency and interpretable coefficients. As mistakes between adjacent age intervals should be penalized less than errors between more distant intervals, an ordinal logistic regression model (OLR) is trained for the age category variable as well. Still, logistic regression often struggles to capture more complex relationships between

---

<sup>7</sup>Map created with <https://www.mapchart.net/europe-detailed.html>

dependent and independent variables. For more performant results, tree ensemble methods are put to use. These techniques are more robust to outliers and are better at discovering non-linear relationships between variables. However, they are computationally intensive and black box. **Random Forest** (RF) uses a bagging approach to combine the votes of multiple decision trees trained on different subsets of the features. Next, three boosting models are considered. **XGBoost** (XGB) is a boosting tree ensemble model that achieves state-of-the-art performance on supervised tasks with tabular data. **Light Gradient Boosting Machine** (LGBM) is designed to train fast, efficiently, and with less memory, while **Catboost** (CB) has built-in support for categorical features.

For all classifiers, hyperparameter tuning is performed using Bayesian Search with 5-fold cross validation. For some classifiers, class weights are added to the objective function to counter class imbalance.

The classifiers are trained on a feature matrix built from raw tweet content and user profile data. The objective is to use information sources that have not been previously processed by LFs, to avoid overfitting, and that can help predict the remaining unlabeled users. Two sets of BoW features are constructed for the profile description and the aggregated tweet corpus per user, respectively. Each BoW set consists of the 500 most frequent unigrams in the training set accounts, with a minimum document frequency of 10 for the profile description and 285 for the tweet corpus. Unigrams include words, emojis, hashtags, and user mentions. As a preprocessing step, all text is converted to lower case and all punctuation marks, numeric characters, stopwords, hyperlinks, and keywords used as part of LFs are removed. A set of 128 topic features is extracted from the aggregated tweet corpus using Top2Vec<sup>8</sup> (Angelov, 2020). Compared to other topic models, Top2Vec has the advantage to not require a predefined number of topics as hyperparameter. Topics are represented as binary variables with one topic assigned to each user. Following Barberá (2016), a set of 256 celebrity-followers features is created. Celebrities are defined as users from the training set and users mentioned in the tweet corpus, who have between 10.000 and 200.000 followers, come from Belgium or the Netherlands, and are verified. A verified account status is accredited by Twitter following a strict procedure<sup>9</sup>. For every celebrity, a binary feature is created with value 1 if a user is following the celebrity on Twitter, and 0 otherwise. Two groups of metadata features complete the feature matrix. A first set studies the information of hyperlinks in profile descriptions, e.g. looking for links with a Belgian “.be” or Dutch “.nl” internet domain or identifying users with a hyperlink containing “linkedin.com” or “instagram.com”. However, no information was retrieved by opening said links. A second set of metadata features contains information on the user’s online behavior: the account age in years, the main type of device used to log in, and the most frequent usage day of the week and period of the day. In total, the feature matrix contains 1417 features.

---

<sup>8</sup><https://github.com/ddangelov/Top2Vec>

<sup>9</sup><https://help.twitter.com/en/managing-your-account/twitter-verified-accounts>

## 4.6 Extended generative model

The best classifier for each task is added to the label matrix as a labeling function. While the discriminative model allows to reach full coverage of the train set, it is expected that its errors are mitigated by the other labeling functions, resulting in overall better predictions. For location, the best classifiers of Task A, B, and C are used conjointly. Task A labels users that are foreign or abstains, task B labels non-Flemish but Belgian users or abstains, and task C assigns one of the five Flemish provinces or abstains.

The extended generative model is then trained on the train set in an unsupervised way and the quality of its predictions are evaluated on the hand-labeled test set as a measure of its ability to generalize to unseen Flemish Twitter users.

The following metrics are used to evaluate the demographic prediction algorithms on the hand-labeled test set: accuracy, macro F1-score, Kendall’s  $\tau$  and Spearman’s  $r_s$ . **Accuracy** (A), the percentage of correctly classified users, is straightforward to interpret. Still, it may be too optimistic in case of strong class imbalance. The **F1-score** expresses the trade-off between Precision and Recall, as formulated in Equation 4. It ranges from 0 to 1, with larger values indicating a better performance. In case of multiclass classification it is more interesting to use the **Macro F1-score** (MF1), where the unweighted mean of all the classes F1-scores is taken. By doing so, all classes are treated equally, regardless of their occurrences.

$$F1 = 2 \frac{Precision \ Recall}{Precision + Recall} \quad (4)$$

where  $Precision = \frac{TP}{TP+FP}$  and  $Recall = \frac{TP}{TP+FN}$  with TP true positive rate, FP false positive rate, TN true negative rate, and FN false negative rate.

In addition, the **Kendall’s  $\tau$**  and **Spearman’s  $r_s$**  coefficients are used to evaluate the age category task. Both metrics are well-suited for ordinal classification tasks as they measure the extent two ranked variables are monotonously related, on a scale of -1 to 1 with a large absolute value indicating a strong correlation. Kendall’s  $\tau$  coefficient is expressed as:

$$\tau = \frac{concordant \ pairs - discordant \ pairs}{concordant \ pairs + discordant \ pairs} \quad (5)$$

where a concordant pair is defined as a pair of observations where the predicted and actual values maintain a similar order and a discordant pair is a pair where these values follow a different order. On the other hand, Spearman’s  $r_s$  is defined as the Pearson correlation coefficient  $\rho$  between the rank variables:

$$r_s = \rho_{R(X),R(Y)} = \frac{cov(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}} \quad (6)$$

where for a sample size  $n$ , raw scores  $X_i$  and  $Y_i$  are converted to ranks  $R$ , with  $cov(R(X), R(Y))$  their covariance and  $\sigma_{R(.)}$  their standard deviation.

## 5 Results and discussion

### 5.1 Weakly labeled set

Applying the base generative label models on the unlabeled training set produces 22.349, 1.249, and 14.481 weak labels for gender, age category, and location respectively. This corresponds to a coverage, i.e. the fraction of the dataset that has at least one label, of 81.5%, 4.5%, and 52.8% respectively. Table 4 displays a decomposition of the coverage per category of LFs. In addition, it shows how different LFs categories overlap with each other. The overlap of a LFs category  $\Lambda \in \mathbf{L}$  is the percentage of data points that are labeled by at least one other LFs category. It is calculated as in Equation 7:

$$overlap(\Lambda) = coverage(\Lambda) + coverage(\mathbf{L} \setminus \{\Lambda\}) - coverage(\mathbf{L}) \quad (7)$$

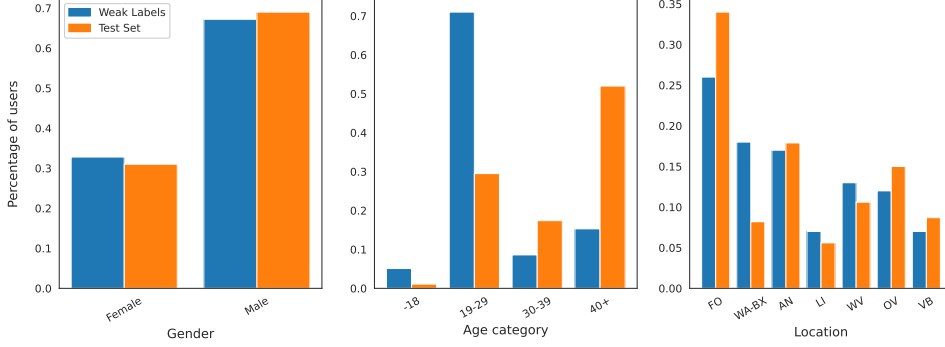
Logically, it holds that  $overlap(\Lambda) \leq coverage(\Lambda)$  for all  $\Lambda \in \mathbf{L}$ .

Gender			Age category			Location		
$\Lambda$	Coverage	Overlap	$\Lambda$	Coverage	Overlap	$\Lambda$	Coverage	Overlap
Keywords	0.076	0.069	Keywords	0.014	0.002	Town names (Location)	0.458	0.096
Name Dictionary	0.529	0.419	RegEx	0.033	0.002	Town names (Description)	0.16	0.094
VGG-Face	0.390	0.383				ZIP	0.008	0.005
CLIP	0.671	0.542						

**Table 4** Percentage of coverage and external overlap of LFs categories  $\Lambda$  on the training set.

For **gender**, the LFs collectively reach the largest portion of the training set. As VGG-Face requires face detection before being able to predict the gender, it withholds more often than CLIP, which assigns labels to all available images. In line with our assumptions, neither of these models is able to reach full coverage as the profile picture can be absent or can show an object. When taking a closer look at overlap, it appears that VGG-Face and keywords introduce few users that were not yet covered by another LF. Contrastingly, the names dictionary and CLIP LFs are responsible for labeling more than 10% of the users. The coverage of the LFs for the **age** category is rather low, but not unexpected as users are less prone to disclose clues about their age. On the contrary, half of the training set users disclosed a valid **location** on their user profile. Most locations were retrieved from the designated location profile field. Some locations were also acquired from the description field, and the overlap is low enough to justify applying LFs on both those profile fields. With a unique coverage of 0.3%, ZIP LFs are significantly less performant. Internal overlaps within a LFs category are shown in Appendix C.

Since the weak labels are not sampled randomly from the training set of unlabeled users, the distribution of the target variable between the weak labels and the test set could differ significantly. This problem, known as prior probability shift (Moreno-Torres et al., 2012), is a form of sample selection bias and can hurt the performance of any predictive model applied on this weakly labeled set. Figure 10 plots the distributions of the weakly labeled set and the test set side by side. The exact percentage of users per demographic category in the weakly labeled and test sets are shown in Appendix D. Unlike gender, age category and location suffer from prior probability



**Fig. 10** Distribution of users according to weakly labeled set and randomly sampled test set.

shift. For age, we observe a clear over-representation of users in their twenties. One explanation for this could be that users in this age category are more keen to disclose their age in their profile description. Though we acknowledge that the potential introduction of a prior probability shift is a significant pitfall of PWS, we leave the assessment and correction of this problem for future work.

## 5.2 Discriminative models

Table 5 discloses the performance of the discriminative models, as evaluated with 5-fold cross-validation on the weakly labeled sets. No large performance gap seems to exist between the various classifiers. The gradient boosting trees are particularly effective on tasks with large weakly labeled sets. More concerning, the performance of all classifiers on the two rank-order metrics ( $\tau$  and  $r_s$ ) for age category is relatively low.

Classifier	Gender		Age category				Location A		Location B		Location C	
	A	MF1	A	MF1	$\tau$	$r_s$	A	MF1	A	MF1	A	MF1
LR	0.69	0.54	0.62	<b>0.39</b>	0.26	0.28	0.87	0.83	0.85	0.72	0.5	0.48
OLR	/	/	0.56	0.31	0.31	0.33	/	/	/	/	/	/
RF	0.72	0.69	<b>0.64</b>	<b>0.39</b>	<b>0.33</b>	0.35	0.88	0.84	0.85	0.65	<b>0.57</b>	<b>0.54</b>
XGB	0.72	<b>0.7</b>	0.59	<b>0.39</b>	<b>0.33</b>	<b>0.37</b>	0.88	0.85	<b>0.88</b>	0.77	0.56	<b>0.54</b>
LGBM	<b>0.75</b>	0.67	0.58	0.37	0.3	0.33	<b>0.89</b>	<b>0.86</b>	<b>0.88</b>	<b>0.78</b>	<b>0.57</b>	<b>0.54</b>
CB	<b>0.75</b>	0.66	0.63	<b>0.39</b>	0.32	0.35	<b>0.89</b>	0.85	0.87	0.75	0.56	0.53

**Table 5** Discriminative models performance with 5-fold cross validation. A is the accuracy, MF1 is the macro F1-score,  $\tau$  is the Kendall’s rank correlation coefficient, and  $r_s$  is the Spearman’s rank correlation coefficient.

For every task, the best classifier is selected as the one with the highest macro F1-score. Ties are resolved by picking the highest accuracy score. To avoid any overfitting, the performance on the hand-labeled test set is not taken into account. Accordingly, XGB is selected for gender, RF for age category and LGBM for location tasks A, B and C. In addition, an analysis of the feature importance is provided in Table 6. Features are grouped in five sets and their importance is computed for the best discriminative model

of each task, using the SHAP tree explainer (Lundberg et al., 2020). The aforementioned method is an implementation of the game theoretic concept of Shapley values. Each cell represents the marginal contribution of a feature set to the corresponding target value and all row values sum to one. For binary tasks, the contribution is the same for both targets, thus only one target is shown in the table. An analysis of the features most associated with each demographic category is shown in Appendix E.

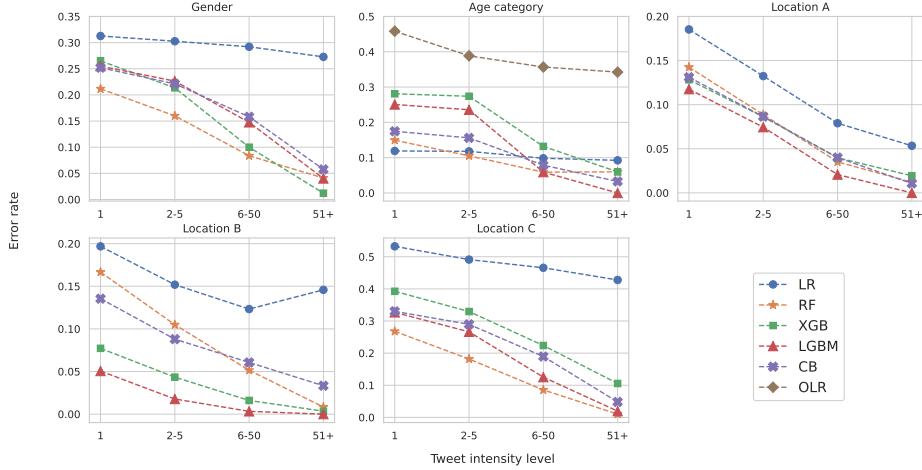
Class	BoW Profile	BoW Tweet	Feature Sets		
			Topic	Celebrity-Followers	Metadata
Female	0.112	0.35	0.011	<b>0.409</b>	0.118
-18	0.083	0.3	0.057	0.182	<b>0.376</b>
19-29	0.07	0.284	0.036	0.232	<b>0.378</b>
30-39	0.049	0.209	0.028	0.281	<b>0.433</b>
40+	0.055	0.305	0.036	0.274	<b>0.331</b>
FO	0.06	0.24	0.014	<b>0.563</b>	0.124
WA-BX	0.065	0.298	0.031	<b>0.529</b>	0.077
AN	0.065	0.241	0.024	<b>0.65</b>	0.019
LI	0.076	0.206	0.032	<b>0.579</b>	0.107
WV	0.045	0.288	0.022	<b>0.596</b>	0.049
OV	0.061	0.268	0.039	<b>0.606</b>	0.026
VB	0.074	0.289	0.064	<b>0.514</b>	0.059

**Table 6** Feature importance by feature groups, as measured by the SHAP tree explainer (Lundberg et al., 2020).

Results clearly indicate that celebrity-followers features are the most important across all gender and location categories, while metadata dominates the age category task. Further investigation into these feature sets shows that, in particular, account age is an important factor for deriving the user’s age, while local politicians and football clubs are highly predictive of user location. While Barberá (2016) did not introduce them for that purpose, we observe here that celebrity-followers features are well-suited for location prediction. In contrast, Topic and BoW profile features barely contribute to the classifiers’ decisions. This could be a sign that heuristic keywords, removed from the description before computing BoW features, already capture all the available demographic information. The BoW tweet features are important across all tasks. However, as previously shown in Figure 7, most users have written less than 5 tweets over the period 2019-2020. This reliance of the classifiers on the BoW tweet features could indicate that classification errors are related with low tweet intensity. Figure 11 confirms this assumption. With few exceptions, the error rate is lower in subsets of users with a higher tweet intensity, across all tasks. In addition, for each task, at least one classifier achieves a near 0% error rate for users with 51 tweets or more. Accordingly, adding more tweets from the user timeline should increase the models’ performance. However, this comes at the cost of more Twitter API queries, more storage space, and more BoW preprocessing time.

### 5.3 Extended generative model

Table 7 shows the performance of three-step PWS, i.e. of the predictions made by the extended generative model, on the hand-labeled test set. We include some benchmarks



**Fig. 11** Error rate per level of tweet intensity.

as well. Firstly, we consider the performance obtained by using the discriminative model as an end-model, which corresponds to the traditional PWS methodology. It is also close to standard weak supervision approaches in demographic prediction (Al Zamal et al., 2012; Barberá, 2016; Culotta et al., 2016; Graells-Garrido et al., 2020), although the weak labels have been generated from multiple sources in our case. Secondly, to assess the usefulness of the generative process, we study the performance of a simple majority vote applied on the extended label matrix. Thirdly, we benchmark our results for gender and age category prediction against those of M3 (Wang et al., 2019) which is the state-of-the-art multimodal and multilingual deep learning model for those two tasks. It takes the profile image, user name, screen name, and profile description of a Twitter profile in raw format as input, and makes predictions using a neural network architecture. M3 was trained on a large dataset, consisting of millions of European user profiles speaking 32 languages, including Dutch. The results obtained by predicting the mode, “Male” for Gender, “40+” for age category, and “Other” for location, are shown as a simple baseline.

Model	Gender		Age category				Location	
	A	MF1	A	MF1	$\tau$	$r_s$	A	MF1
Three-step PWS	0.916	0.903	<b>0.555</b>	<b>0.411</b>	0.443	0.487	<b>0.738</b>	<b>0.688</b>
Discriminative Model	0.742	0.715	0.533	0.337	0.385	0.42	0.622	0.544
Extended Majority Vote	0.86	0.579	0.533	0.29	0.354	0.396	0.499	0.515
M3	<b>0.922</b>	<b>0.904</b>	<b>0.555</b>	0.373	<b>0.5</b>	<b>0.558</b>	/	/
Mode	0.689	0.408	0.519	0.171	/	/	0.341	0.073

**Table 7** Final classification results. A is the accuracy, MF1 is the macro F1-score,  $\tau$  is the Kendall’s rank correlation coefficient, and  $r_s$  is the Spearman’s rank correlation coefficient.

For all tasks, three-step PWS surpasses the traditional PWS methodology, as it outperforms the discriminative model as an end-model. On location prediction, the extended generative model attains an accuracy of 73.8% and a macro F1-score of 68.8%, which is

considerable for a 7-class classification problem. In addition, it exceeds the performance of the extended majority vote, highlighting the usefulness of using an unsupervised generating process to combine the weak labels. The performance gap with both benchmarks is smaller for the prediction of age category, which is due to the lower coverage, overlap, and amount of interdependencies within the age category labeling matrix. Furthermore, the extended generative model achieves comparable performance with M3, while being trained on a much smaller dataset. For age category, it even improves the macro F1-score with 3.8 percent points. On the contrary, it lacks performance with regard to the two ordinal metrics. While comparing, it is important to keep in mind that M3 was trained with more data. We compare the predictions of three-step PWS and M3 using the McNemar’s test, a frequently used significance test for comparing ML classifiers (Dietterich, 1998). We observe no statistically significant difference between three-step PWS and M3 for gender classification (p-value=0.5114). However, we observe a statistically significant difference between three-step PWS and M3 for age category classification (p-value=0.008). Confusion matrices of three-step PWS predictions are reported in Appendix F.

The total training time of three-step PWS can be broken down as follows. Generating the train and test label matrices takes 77, 16, and 300 seconds for gender, age category, and location prediction, respectively. Training the base and extended models with Snorkel takes less than a second. The average training time of the discriminative models varies from one classifier to another. It ranges from less than 60 seconds for logistic regression models, to more than 300 seconds for boosting trees.

## 5.4 Three-step PWS on new target populations

In this section, we briefly address how to apply the experiments to other areas of interest and demographic categories and highlight the flexibility of three-step PWS with regard to this.

A change in the demographic categories, e.g. splitting the *40 and above* category into *40-59* and *60 and above*, requires two modifications. First, the targets need to be redefined. Second, keywords, distant supervision sources, and other LFs, needs to be updated for the new demographic categories. From there, the base generative, discriminative, and extended generative model are automatically updated. Because we leverage logistic regression and tree ensembles as discriminative classifiers, the update takes a matter of minutes.

Studying another geographic area of interest, e.g. France or Spain, requires the same two modifications, except that all location categories are redefined. Weak supervision sources such as keywords or town lists are matched with their corresponding demographic category. For gender and age, new keywords and distant supervision sources might need to be collected and new LFs defined to match the local language.

Thanks to the flexibility of three-step PWS, the freedom is given to the researcher to decide which LFs to include in the base generative model, and how to define demographic categories. We believe this framework will help researchers and National Statistical Offices understand better the demographic patterns of the Twitter population.

## 6 Conclusion

### 6.1 Main contributions

This work framed social media demographic prediction as a PWS problem, and introduced three-step PWS, a new methodology to predict the gender, age category, and location of Twitter users. Our results show that three-step PWS improved the results obtained with a traditional PWS methodology and is competitive with the state-of-the-art pretrained model, while being trained on a much smaller dataset, in a short amount of time, and offering additional flexibility. This makes particularly suitable to perform demographic inference on small target population, e.g. Flanders. In addition, three-step PWS does not require expensive hand-labeling to create a training set. Furthermore, we proposed a set of labeling functions suitable for demographic prediction. Given their high accuracy and coverage, third-party models are the most interesting weak supervision source. Distant supervision sources, when properly collected, become efficient and precise labeling functions. As for keywords and regular expression based heuristics, their low coverage does not compensate their time-consuming design. Based on these results, we recommend to prioritize the use of relevant third-party models and distant supervision sources. If no such LFs is available, e.g. for age prediction, keywords and regular expressions should be considered.

### 6.2 Limitations and future work

Unfortunately, the proposed methodology has its limitations. First, the weakly labeled set may suffer from a strong prior probability shift, as was the case for the age category task. To the best of our knowledge, a solution is yet to be found to solve this problem. Secondly, we restricted our study to geolocated users which constitute only a subset of the whole Twitter population. Whether this impacts the performance of three-step PWS is an interesting direction for future research. Thirdly, our study considered a very restricted range of demographic values. Future research could expand the current PWS process by adding more categories to the gender task, by predicting more fine-grained age categories and locations, or by including other demographic attributes like education and income level. Thanks to the flexibility of the PWS framework, those extensions and modification can be easily incorporated, unlike with pre-trained deep learning models like M3. Eventually, our model does not integrate a company account detection component, as it relies on M3 predictions instead, and it ignores the potential presence of bots and fake accounts (Alarifi et al., 2016). Future research could include those two subtasks in the three-step PWS methodology.

Leveraging the demographics of users is critical to many applications relying on Twitter data, for example when the objective is to draw conclusions on the general population. Our methodology can provide the necessary input for demographic debiasing models like post-stratification (Wang et al., 2019) and resampling (Wang et al., 2020). An interesting future research direction is to design new end-to-end workflows that optimally combine demographic prediction and debiasing models.

Recent advances made in the PWS field could also be considered for future research. Instead of hand-writing LFs for all target variables, it is possible to generate them

automatically, or in an active learning process (Zhang et al., 2022). A richer labeling model can be obtained by using Partial LFs (Yu et al., 2022), which return a set of potential labels instead of one. This is useful when some characteristics are shared by a subset of the possible labels.

### 6.3 Ethical aspects

All the data collected for our experiments was processed in compliance with the rules of Twitter Academic Search API as they were in force in spring 2022. To respect users' data privacy, coarse classes are applied to the demographic attributes, i.e. we use age brackets instead of exact value for age, and provinces instead of exact coordinates for location. For the same reason, the output of our model is only meant to be used in an aggregated way, e.g. as background demographic information for official statistics based on social media indicators. We acknowledge the potential presence of fairness issues in the demographic prediction process and the need to tackle it in future work. While the label matrix and the generative model's weights allows us to interpret to some extent the decision process of three-step PWS, more work is still needed to make PWS models fully interpretable. Only our code and a small dataset of fictional users are publicly made available, without the Twitter data used in this study. The code allows researchers, who obtained credentials after a positive review of their research proposal from Twitter, to conduct similar experiments on other target populations.

## 7 Acknowledgments

This research was funded by the Statistics Flanders research cooperation agreement on Data Science for Official Statistics.

### Appendix A List of abbreviations

Abbreviations used in the paper are listed in table [A1](#).

### Appendix B List of Keywords for Heuristics LFs

Table [B2](#) lists all keywords used in the heuristic labeling functions for Gender and Age category prediction. It is a combination of Dutch and English keywords.

Related Work	BHM	Bayesian hierarchical model
	BoW	Bag of words
	LLP	Learning from labeled proportions
	RLM	Regularized linear models
	SVM	Support vector machines
	TF-IDF	Term frequency-inverse document frequency
Locations	AN	Antwerpen (Antwerp)
	FO	Foreign (non-Belgian)
	LI	Limburg
	OV	Oost-Vlaanderen (East Flanders)
	VB	Vlaams-Brabant (Flemish Brabant)
	WA-BX	Wallonia and Brussels
	WV	West-Vlaanderen (West Flanders)
Methodology	CB	CatBoost
	CLIP	Contrastive language-image pre-training
	LF	Labeling functions
	LGBM	Light Gradient boosting machine
	LR	Logistic regression
	OLR	Ordinal logistic regression
	PWS	Programmatic weak supervision
	RF	Random forest
	XGB	Extreme gradient boosting
Metrics	A	Accuracy
	MF1	Macro F1-score
	$\tau$	Kendall’s rank correlation coefficient
	$r_s$	Spearman’s rank correlation coefficient

**Table A1** List of abbreviations

## Appendix C Labeling functions internal overlaps and conflicts

Table C3 complements Table 4 by showing the overlaps and conflicts within a LFs category  $\Lambda$ , in percentage of users in the training set. A conflict occurs when at least two LFs do not abstain and disagree on the label they assign to a particular instance.

## Appendix D Percentage of users per demographic category

Table D4 complements Figure 10 with the percentages of users per demographic category in the weakly labeled set and hand-labeled test set.

Gender		Age category			
Male	Female	-18	19-29	30-39	40+
he/him	she/her	middelbaar	twintiger	dertiger	veertiger
hij/hem	zij/haar	lyceum	twenty-something	thirty-something	fourty-something
hij	zij	leerling	20something	30something	40something
he/they	zij/hun		20-something	30-something	40-something
hij/hun	she/they		student	thirty	50something
hem	girly		studente		50-something
dad(dy)	mom(my)				60something
father	mother				60-something
vader	moeder				grandmother
papa	mama				oma
man	vrouw				grootmoeder
dude	meisje				grandfather
brother	sister				opa
broer	zus				grootvader
uncle	aunt				nonno
nonkel	tante				pensioen
oom	mum				gepensioneerd
kerel	studente				gepens.
homo	woman				retired
boerenzoon	queen				emeritus
zoon	sis				kleinkinderen
	feministe				petits-enfants
	keizerin				grandchildren
	eigenares				

**Table B2** Keywords used for the gender and age category labeling functions

A	Gender		A	Age category		A	Location	
	Internal Overlap	Internal Conflict		Internal Overlap	Internal Conflict		Internal Overlap	Internal Conflict
Keywords	0.001	0.001	Keywords	0	0	Town names (Location)	0.018	0.017
Name Dictionary	0	0	RegEx	0.003	0.0003	Town names (Description)	0.031	0.031
VGG-Face	0	0				ZIP	0	0
CLIP	0	0						

**Table C3** Percentage of internal overlaps and conflicts for all LFs categories A applied on the training set.

Demographic value	Training set	Test set
Female	0.328	0.31
Male	0.672	0.69
-18	0.051	0.011
19-29	0.71	0.295
30-39	0.086	0.174
40+	0.153	0.52
FO	0.262	0.34
WA-BX	0.177	0.082
AN	0.169	0.179
LI	0.068	0.056
WV	0.131	0.106
OV	0.117	0.15
VB	0.075	0.087

**Table D4** Percentage of users per demographic category in the weakly labeled set and the test set.

## Appendix E Analysis of the features association with each demographic category

For every feature (BoW, Topic, celebrity-followers, and metadata) we compute its association with a demographic category. This is calculated as the percentage of training set users for which the best discriminative model predicts the corresponding category, excluding users with null feature values. Tables E1, E2, and E3 show the top 10 features for every demographic category. Note that this does not imply that those features are directly used by the classifiers. It only indicates that a large share of the non-null rows of said feature are paired with a specific demographic category, as a result of the classifier learning process.

The gender features in Table E1 show that the XGB model has frequently paired emojis in profile descriptions with the female users, while male profiles are associated with sport channels and vocabulary related to gaming and computer science. These features highlight a biased view on gender and gives precedence to further investigation and potentially a fairness analysis. Identifying these features is also interesting for spotting omitted keyword LFs candidates, for example ‘guy’ and ‘husband’/‘echtgenoot’ for Male.

Demographic category	Features
Female	profile: ✨, profile: 🦋, profile: 💕, profile: ❤️, profile: ♀, profile:fashion, profile: 🌻, profile: 🧑, profile:lezen, profile: ❤️
Male	profile:cloud, profile:guy, follow:@melindafarrell, profile:echtgenoot, profile:software, profile:gamer, follow:@ElevenSportsBEn, follow:@KVCWesterlo, profile:husband, profile:developer, follow:@ElevenSportsBEf

Fig. E1 Gender features

The age associated features are displayed in Table E2. Unfortunately, for most age categories there is no clear trend in the features, except for the 40+ age category. It characterizes itself, according to the classifier, by mentioning Belgian politicians or political parties in tweets.

Demographic category	Features
-18	profile:pro, follow:@SBS6, profile:years, profile:first, topic_49:tourism, profile:🤩, topic_94:DeTijdLoze, profile:samen, profile:iedereen, has:twitch_url
19-29	profile:⭐, tweet:🤔, follow:@melindafarrell, profile:🌟, tweet:hahaha, tweet:wa, follow:@kastiop, tweet:ni, tweet:brussels, profile:🦋
30-39	has:discord_url, follow:@kallenje, follow:@brittije, follow:@NetflixBeNL, follow:@MobileVikingsBE, profile:🤩, profile:👩, follow:@RealKateRyan, follow:@ransbottyn, profile:marketeer
40 +	tweet:@torfsrik, tweet:partijen, tweet:@groen, tweet:@kristofcalvo, tweet:@vlbelang, tweet:@phroose, tweet:@cdenv, tweet:@spa, tweet:@jdeceulaer, tweet:@bartdewever

**Fig. E2** Age features

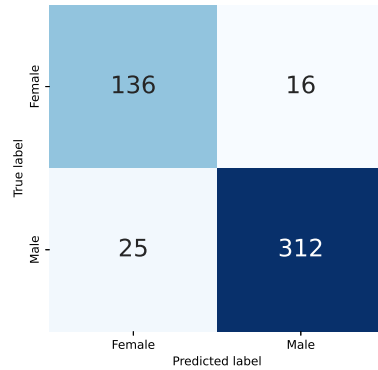
Table E3 shows the top features for the location categories. Celebrity-followers features are dominating, notably the verified accounts of football clubs, local authorities, and universities are associated with their respective province. The top 10 features of the FO and WA-BX categories are all semantically related to the Netherlands and Brussels, respectively, highlighting they constitute the largest subgroup in their demographic category.

Demographic category	Features
FO	has:NL_domain, profile:titel, profile:feeynoord, profile:ajax, follow:ZS_racing, follow:@brittije, follow:@24/7BZ, follow:@heelhollandbakt, profile:por, profile:radslid
WA-BX	follow:@jackeparrock, follow:@fadilalaanan, follow:@lapremiere, follow:@irelandrepbru, follow:@mvanhulten, follow:@STIBMIVB, follow:@lecho, follow:@sophieintVeld, profile:policy, profile:european
AN	profile:antwerp, follow:@Stad_Antwerpen, profile:🇳🇱, follow:@PZAntwerpen, Topic_124:RobTV, follow:@atvbe, profile:feeynoord, follow:@BZAntwerpen, profile:singer
LI	follow:@kallenje, follow:@vuka20, follow:@kRCGenkOfficial, follow:@ZS_racing, follow:@BenjAlvarez1, follow:@24/7BZ, profile:certified, follow:@pukkelpop, profile:waar, follow:@STTV
WV	follow:@FocusWTV, tweet:kortrijk, tweet:westvlaanderen, follow:@BMechele44, follow:@kvkofficieel, follow:@ClubBrugge, follow:@barco, tweet:brugge, follow:@SimonsTimmy, follow:@cercleofficial
OV	follow:@oost_vlaanderen, follow:@Stadgent, profile:ghent, profile:@ugent, follow:@KAAGent, tweet:oostvlander, profile:del, follow:@UGent, profile:msc, follow:@vooruit
VB	follow:@PolitieLeuven, profile:@kuleuven, tweet:leuven, follow:@Jujuca1987, follow:@KULeuven, follow:@OHLeuven, follow:@irelandrepbru, follow:@BRUZZbe, follow:@jackeparrock

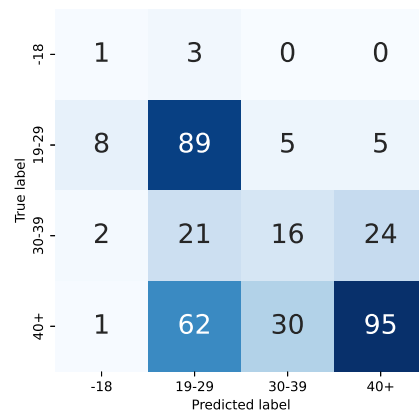
**Fig. E3** Location features

## Appendix F Confusion matrices of the extended generative model

Figures F4, F5, and F6 complete the results of Table 7 by providing the confusion matrices for the extended generative model.



**Fig. F4** Gender prediction Confusion Matrix on the test set



**Fig. F5** Age category prediction Confusion Matrix on the test set

## References

- Barberá, P.: Less is more? How demographic sample weights can improve public opinion estimates based on Twitter data. Working Paper NYU (2016)
- Hou, W., Li, Y., Liu, Y., Li, Q.: Leveraging multidimensional features for policy opinion sentiment prediction. Information Sciences (2022)

True label	Predicted label						
	FO	WA-BX	AN	LI	WV	OV	VB
FO	110	13	3	8	3	4	0
WA-BX	1	29	1	0	1	1	1
AN	0	2	49	3	2	4	14
LI	0	0	1	11	1	1	9
WV	0	0	2	2	35	1	4
OV	3	5	2	2	4	43	2
VB	0	0	3	2	1	2	28

**Fig. F6** Location prediction Confusion Matrix on the test set

Culotta, A.: Reducing sampling bias in social media data for county health inference. In: Joint Statistical Meetings Proceedings, pp. 1–12 (2014). Citeseer

Diaz, F., Gamon, M., Hofman, J.M., Kıcıman, E., Rothschild, D.: Online and social media data as an imperfect continuous panel survey. *PloS one* **11**(1) (2016)

Vijayaraghavan, P., Vosoughi, S., Roy, D.: Twitter demographic classification using deep multi-modal multi-task learning. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 478–483. Association for Computational Linguistics, Vancouver, Canada (2017). <https://doi.org/10.18653/v1/P17-2076> . <https://aclanthology.org/P17-2076>

Biffignandi, S., Bianchi, A., Salvatore, C.: Can big data provide good quality statistics? A case study on sentiment analysis on Twitter data. In: Int. Total Surv. Error Workshop ITSEW-2018 DISM-Duke Initiat. Surv. Methodol (2018)

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., Rosenquist, J.: Understanding the demographics of Twitter users. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 5 (2011)

Vandendriessche, K., Steenberghe, E., Matheve, A., Georges, A., De Marez, L.: imec.digimeter 2020, Digitale trends in Vlaanderen (2020). <https://www.imec.be/sites/default/files/inline-files/DIGIMETER2020.pdf>

Ratner, A.J., De Sa, C.M., Wu, S., Selsam, D., Ré, C.: Data programming: Creating large training sets, quickly. *Advances in neural information processing systems* **29**

(2016)

- Zhang, J., Hsieh, C.-Y., Yu, Y., Zhang, C., Ratner, A.: A survey on programmatic weak supervision. arXiv preprint arXiv:2202.05433 (2022)
- Wang, Z., Hale, S., Adelani, D.I., Grabowicz, P., Hartman, T., Flöck, F., Jurgens, D.: Demographic inference and representative population estimates from multilingual social media data. In: The World Wide Web Conference, pp. 2056–2067 (2019)
- Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021). PMLR
- Rao, D., Paul, M., Fink, C., Yarowsky, D., Oates, T., Coppersmith, G.: Hierarchical bayesian models for latent attribute detection in social media. In: Fifth International AAAI Conference on Weblogs and Social Media (2011)
- Matz, S.C., Menges, J.I., Stillwell, D.J., Schwartz, H.A.: Predicting individual-level income from Facebook profiles. PLOS ONE **14**(3), 1–13 (2019) <https://doi.org/10.1371/journal.pone.0214369>
- Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T.: “How old do you think I am?” A study of language and age in Twitter. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 7 (2013)
- Preoțiu-Pietro, D., Volkova, S., Lampos, V., Bachrach, Y., Aletras, N.: Studying user income through language, behaviour and affect in social media. PLOS ONE **10**(9), 1–17 (2015) <https://doi.org/10.1371/journal.pone.0138717>
- Al Zamal, F., Liu, W., Ruths, D.: 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, vol. 6, pp. 387–390 (2012)
- Ikeda, K., Hattori, G., Ono, C., Asoh, H., Higashino, T.: Twitter user profiling based on text and community mining for market analysis. Knowledge-Based Systems **51**, 35–47 (2013)
- Miranda Filho, R., Almeida, J.M., Pappa, G.L.: Twitter population sample bias and its impact on predictive outcomes: A case study on elections. In: 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1254–1261 (2015). IEEE
- Culotta, A., Ravi, N.K., Cutler, J.: Predicting Twitter user demographics using distant supervision from website traffic data. Journal of Artificial Intelligence Research **55**, 389–408 (2016)

- Graells-Garrido, E., Baeza-Yates, R., Lalmas, M.: Representativeness of abortion legislation debate on Twitter: A case study in Argentina and Chile. In: Companion Proceedings of the Web Conference 2020, pp. 765–774 (2020)
- Compton, R., Jurgens, D., Allen, D.: Geotagging one hundred million Twitter accounts with total variation minimization. 2014 IEEE International Conference on Big Data, IEEE Big Data 2014, 393–401 (2014) <https://doi.org/10.1109/BigData.2014.7004256>
- Jurgens, D., Finethy, T., McCorriston, J., Xu, Y.T., Ruths, D.: Geolocation prediction in Twitter using social networks: A critical analysis and review of current practice. In: Ninth International AAAI Conference on Web and Social Media (2015)
- Rahimi, A., Cohn, T., Baldwin, T.: Semi-supervised user geolocation via graph convolutional networks. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2009–2019 (2018)
- Hinds, J., Joinson, A.N.: What demographic attributes do our digital footprints reveal? A systematic review. PloS one **13**(11), 0207112 (2018)
- HaCohen-Kerner, Y.: Survey on profiling age and gender of text authors. Expert Systems with Applications, 117140 (2022)
- Pennacchiotti, M., Popescu, A.-M.: A machine learning approach to Twitter user classification. In: Fifth International AAAI Conference on Weblogs and Social Media (2011)
- Li, J., Ritter, A., Hovy, E.: Weakly supervised user profile extraction from Twitter. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 165–174 (2014)
- Chen, X., Wang, Y., Agichtein, E., Wang, F.: A comparative study of demographic attribute inference in Twitter. Proceedings of the International AAAI Conference on Web and Social Media **9**(1), 590–593 (2015)
- Preoțiuc-Pietro, D., Lampos, V., Aletras, N.: An analysis of the user occupational class through Twitter content. 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), pp. 1754–1764 (2015)
- Ardehaly, E.M., Culotta, A.: Co-training for demographic classification using deep learning from label proportions. In: 2017 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 1017–1024 (2017). <https://doi.org/10.1109/ICDMW.2017.144>
- Aletras, N., Chamberlain, B.P.: Predicting Twitter user socioeconomic attributes with network and language information. In: Proceedings of the 29th on Hypertext and

- Social Media, pp. 20–24 (2018)
- Pan, J., Bhardwaj, R., Lu, W., Chieu, H.L., Pan, X., Puay, N.Y.: Twitter homophily: Network based prediction of user’s occupation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2633–2638 (2019)
- López-Monroy, A.P., Gonzalez, F.A., Solorio, T.: Early author profiling on Twitter using profile features with multi-resolution. *Expert Systems with Applications* **140**, 112909 (2020)
- Wood-Doughty, Z., Xu, P., Liu, X., Dredze, M.: Using noisy self-reports to predict Twitter user demographics. In: Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media, pp. 123–137 (2021)
- Suman, C., Naman, A., Saha, S., Bhattacharyya, P.: A multimodal author profiling system for tweets. *IEEE Transactions on Computational Social Systems* **8**(6), 1407–1416 (2021)
- Angelov, D.: Top2Vec: Distributed Representations of Topics. arXiv (2020). <https://doi.org/10.48550/ARXIV.2008.09470> . <https://arxiv.org/abs/2008.09470>
- McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. *Annual review of sociology* **27**(1), 415–444 (2001)
- Rao, D., Yarowsky, D., Shreevats, A., Gupta, M.: Classifying latent user attributes in Twitter. In: Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents, pp. 37–44 (2010)
- Ardehaly, E.M., Culotta, A.: Mining the demographics of political sentiment from twitter using learning from label proportions. In: 2017 IEEE International Conference on Data Mining (ICDM), pp. 733–738 (2017). <https://doi.org/10.1109/ICDM.2017.84>
- Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., Lazer, D.: Fake news on Twitter during the 2016 US presidential election. *Science* **363**(6425), 374–378 (2019)
- Daas, P.J., Burger, J., Le, Q., Bosch, O., Puts, M.: Profiling of Twitter Users: a Big Data Selectivity Study, pp. 1–25 (2016)
- Ratner, A., Bach, S.H., Ehrenberg, H., Fries, J., Wu, S., Ré, C.: Snorkel: Rapid training data creation with weak supervision. In: Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases, vol. 11, pp. 269–282 (2017). NIH Public Access
- Ratner, A., Hancock, B., Dunnmon, J., Sala, F., Pandey, S., Ré, C.: Training complex models with multi-task weak supervision. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 4763–4771 (2019)

- Fu, D., Chen, M., Sala, F., Hooper, S., Fatahalian, K., Ré, C.: Fast and three-rious: Speeding up weak supervision with triplet methods. In: International Conference on Machine Learning, pp. 3280–3291 (2020). PMLR
- Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: Proceedings of the British Machine Vision Conference (BMVC), pp. 41–14112 (2015). <https://doi.org/10.5244/C.29.41> . <https://dx.doi.org/10.5244/C.29.41>
- Serengil, S.I., Ozpinar, A.: LightFace: A hybrid deep face recognition framework. In: 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), pp. 23–27 (2020). <https://doi.org/10.1109/ASYU50717.2020.9259802> . IEEE. <https://doi.org/10.1109/ASYU50717.2020.9259802>
- Serengil, S.I., Ozpinar, A.: Hyperextended LightFace: A facial attribute analysis framework. In: 2021 International Conference on Engineering and Emerging Technologies (ICEET), pp. 1–4 (2021). <https://doi.org/10.1109/ICEET53442.2021.9659697> . IEEE. <https://doi.org/10.1109/ICEET53442.2021.9659697>
- Moreno-Torres, J.G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N.V., Herrera, F.: A unifying view on dataset shift in classification. *Pattern recognition* **45**(1), 521–530 (2012)
- Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I.: From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence* **2**(1), 56–67 (2020)
- Dietterich, T.G.: Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation* **10**(7), 1895–1923 (1998)
- Alarifi, A., Alsaleh, M., Al-Salman, A.: Twitter turing test: Identifying social machines. *Information Sciences* **372**, 332–346 (2016)
- Wang, Z., Yu, Z., Fan, R., Guo, B.: Correcting biases in online social media data based on target distributions in the physical world. *IEEE Access* **8**, 15256–15264 (2020)
- Yu, P., Ding, T., Bach, S.H.: Learning from multiple noisy partial labelers. In: International Conference on Artificial Intelligence and Statistics, pp. 11072–11095 (2022). PMLR