Assessing Socioeconomic Status of Twitter Users: A Survey

Dhouha Ghazouani[1], **Luigi Lancieri**[2], **Habib Ounelli**[1] and **Chaker Jebari**[3]

[1]**Faculty of Sciences of Tunis, University Campus, Tunis 1060, Tunisia**
(dhouha.ghazouani,habib.ounelli)@fst.utm.tn
[2]**Univ.Lille, Research Center in Signal and Automatic Computing of Lille, F-59000 Lille, France**
luigi.lancieri@univ-lille.fr
[3]**Information Technology Department, Colleges of Applied Sciences, Ibri, Oman**
jebarichaker@yahoo.fr

Abstract

Every day, the emotion and opinion of different people across the world are reflected in the form of short messages using microblogging platforms. Despite the existence of enormous potential introduced by this data source, the Twitter community is still ambiguous and is not fully explored yet. While there are a huge number of studies examining the possibilities of inferring gender and age, there exist hardly researches on socioeconomic status (SES) inference of Twitter users. As socioeconomic status is essential to treating diverse questions linked to human behavior in several fields (sociology, demography, public health, etc.), we conducted a comprehensive literature review of SES studies, inference methods, and metrics. With reference to the research on literature's results, we came to outline the most critical challenges for researchers. To the best of our knowledge, this paper is the first review that introduces the different aspects of SES inference. Indeed, this article provides the benefits for practitioners who aim to process and explore Twitter SES inference.

1 Introduction

The ability to identify the socioeconomic status of social media users accurately is beneficial for the individual scale as well as the societal one. This field starts to be a well-explored research domain. The difficulty to identify the socioeconmic status of authors and the lack of explicit personal information have brought with them some challenge for computer scientists.

Nowadays, Twitter's monthly active members exceed 300 millions. These members generate over 500 million conversations (tweets) daily [1]. These conversations are short text messages including a maximum of 140 characters (recently extended to 280). Indeed, this shortage of characters leads to unstructured and noisy texts to the point that natural language processing (NLP) tools cannot manage successfully (Ritter et al., 2011). Moreover, more deduction is required to detect the underlying features of Twitter users. In this regard, researchers and specialized centers will explore and analyze the available demographic information of Twitter users. The results provided by the Pew Research Center (Smith and Brenner, 2012), a subsidiary of the Pew Charitable Trust, show that the majority of Twitter users in the United States are young, with high educational level and exposing a bigger political interest. Authors concluded that focusing on a community characterized by high level of involvement in societal issues (Li et al., 2015) could be fruitful. In a first study (Preoţiuc-Pietro et al., 2015a) prove that language use in social media is an indicator of user's occupational class. In a second study, (Preoţiuc-Pietro et al., 2015b) provide a comparison between income and psycho-demographic traits of Twitter users; among the results of this research they concluded that the rich users expose less emotional status but more neutral content, expressing anger and fear, but less surprise, sadness, and disgust. Recently, (Flekova et al., 2016) found that the writing style can also indicate the income of the users. The higher income is an indicator of education and conscientiousness. Moreover, (Volkova and Bachrach, 2016) concluded that the highly educated users have a stronger tendency to express less sadness and are likely to show more neutral opinions.

User's socioeconomic status (SES) is the most

[1]https://about.twitter.com/company

Proceedings of Recent Advances in Natural Language Processing, pages 388–398,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_046

important predictors of a person's morbidity and mortality experience (Kitagawa and Hauser, 1973; Marmot et al., 1987). The significant impact of SES on public health renders its definition and measurement of critical importance. When the SES is low, it does not only involves poverty and poor health, but it also affects the educational achievements hence the whole society. Thus, the research of (Morgan et al., 2009) finds that children from low-SES households develop a slow academic behavior than children belonging to higher SES groups. For these reasons, marketing campaigns, as well as economic and sociological studies, have found it interesting to determine the socioeconomic status of particular persons.

This article is going to be divided into six sections. After the introduction, section 2 will discuss the metrics of socioeconomic status used with Twitter data. Section 3 will discuss SES indicators and its features. Section 4 will examine the different techniques employed in SES inference and section 5 will present the data collection and analysis process. Section 6 is going to be the conclusion for this article opening the horizons for further discussions.

2 Evaluation of Socioeconomic Status

Socioeconomic status (SES) can be defined as one's possession of social, financial, cultural, and human capital resources. Parental and neighborhood properties are considered as additional components (Cowan et al., 2012). We can also note that SES is a complex unit of measurement of a person's economic and sociological standing, like for instance, his prestige, power or else his economic well-being (Hoff et al., 2002; Oakes and Rossi, 2003). Consequently, one can conclude that the SES is a complex measure of evaluation that differs from a research to another because it takes into account the work experience, the economic position, or the social status.

The concept of the SES detection in literature goes back to the beginning of the 20^{th} century (Chapman and Sims, 1925). In the 20^{th} century the evaluation of SES was based on questions like: How many years did your father go to school?", Do you have a telephone?" or Do you work out of school hours?". Currently, there is an agreement that SES is influenced by three significant factors: the cultural (comprised of skills, capacities, and knowledge), the social (social network combined with the status and power of the people in that network) and the material capital (Jones et al., 2007). Similarly, the primary metrics of the SES are education, occupation, income levels and wealth or lifestyle (Van Berkel-Van Schaik and Tax, 1990; White, 1982).

People are usually divided into groups according to these metrics, from the least advantaged to the most advantaged, medium, or high SES. **Education** is one of the widely used indicator and it is considered by many to be the canonical element of SES because of its influence on later income and occupation (Krieger et al., 1997). This index can be defined by two dimensions: the field of education and the level at which the education was followed. **Income** reflects spending power, housing, diet, and medical care. **Occupation** measures like prestige, responsibility, physical activity, as well as work exposures. The occupational status influences the social capital of individuals and it strengthen the connection with more professional people enjoying wealth and power. Similarly, education indicates skills requisite for acquiring a positive social, psychological, and economic resources (Antonovsky, 1967). Likewise, social classes are measurements that, like SES, aim to locate ones position in the social hierarchy. classes are social categories sharing subjectively-salient attributes used by people to rank those categories within a system of economic stratification" (Wright and Ritzer, 2003). By refering to the definition presented by (Wright and Ritzer, 2003), classes refer to how people are objectively located in distributions of material inequality".

Before the rise of social networks, different studies have looked into other data sources from various domains, like internet browsing behaviors, written texts, telephone conversations, real-world mobile network and communication records.

- (French, 1959): introduced the relationship between different measures of 232 undergraduate students and their future jobs. This work concluded that occupational membership could be predicted with the use of variables such as the ability of persons in using mathematical and verbal symbols, the social class of family and the personality components.

- (Schmidt and Strauss, 1975): have also designed the relationship between the types of occupation and the particular demographic attributes such as gender, race, experience,

education, and location. Their study identified biases and the types of discrimination that can possibly exist in various types of occupations.

The recent excessive use of online social media and the user-generated content in microblogging platforms such as google+, Facebook, or Sine Weibo [2] has allowed the study of author profiling on an unprecedented scale.

- (Li et al., 2014): proposed a framework for assessing the user's features on Twitter using Google+ API. They constructed a publicly available dataset using distant supervision. They submitted their model on three user profile attributes, i.e., Job, Spouse and Education.

- (Zhong et al., 2015): investigated the predictive power of location check-in, extracted from points of interest of Sina Weibo. In order to determine the demographic attributes of the users such as education background (university and non-university), marital status (single, courtship, in love or married) using human mobility as an informative and fundamental user behavior. They developed a comprehensive location to profile (L2P) framework to detect temporality, spatiality, and location knowledge at the same time.

- (Sullivan et al., 2018) has recently reported that Facebook has patented technology that utilizes a sample decision tree to determine its users' social class. Decision tree uses as an input information about a user's demographic information, device ownership, internet usage, household data, etc. The output provides a probability that the user belongs to a given socioeconomic class: working class, middle class or upper class.

Recent studies tackled the inference of socioeconomic characteristics of Twitter users.

- (Lerman et al., 2016): analyze a large corpus of geo-referenced tweets posted by social media users from US metropolitan areas. They measure emotions expressed in the tweets posted from a particular area with the inference of socioeconomic characteristics. They collect Twitter accounts which users are located in Los Angeles. Concerning the sentiment analysis, they used SentiStrength [3]. The study shows that people with higher incomes are associated with weaker social ties.

- (Quercia et al., 2012): treat the relationship between sentiment expressed in tweets and the community socioeconomic well-being. In their research, they collect Twitter accounts which users are located in London. Concerning the sentiment analysis, they used word count technique and the maximum entropy classifier. Socio-demographic data obtained from Index of Multiple Deprivation scores (composite score based on income, employment, education, health, crime, housing, and the environmental quality for each community) of each of the 78 census areas in London.

3 SES Features and Indicators

The quality of features influences the value of a machine learning pattern from which it originates. Microblogging platforms offer a different number of potential features. Different traditional text-based corpora features are used to explore the relationship between these characteristics. Different types of indicators can help infer the SES of Twitter users used over years. This idea is going to be developed later in the article.

3.1 Message Content

Twitter message text represents the backbone of most research works within the field of SES inference as this helps to understand the context of messages themselves. The messages of the social media platform include abbreviations and non-standard formulation as there is no precise rule of writing since most of the tweets are sent via mobile phones.

In his thesis, (Mentink, 2016) used Bag-of-Words to analyze the discussed topics of users. (Preoţiuc-Pietro et al., 2015b) used clustering algorithms to build a list of most frequent unigrams and then they reached their vector representations, consequently using Word2Vec model to compute dense word vectors (grouping words into clusters or topics). While (Lampos et al., 2016) applied spectral clustering to derive clusters of 1-gram that

[2] http://www.weibo.com/signup/signup.php

[3] http://sentistrength.wlv.ac.uk/

capture some potential topics and linguistic expressions, (Preoţiuc-Pietro et al., 2015a) used Normalized Pointwise Mutual Information (NPMI) to compute word to word similarity, then applied singular value decomposition (SVD) to obtain an embedding of words into a low-dimensional space. A good approach of content analysis would take into consideration all possible instances of SES indicators being expressed within the message. For most studies, the use of message content aims at inferring morphological characteristics and language use.

(Barberá, 2016) used the emoji characters as features (bag-of-emoji) and the author used word counts as another features (bag-of-words) with the application of TF-IDF transformation. In order to obtain a robust result, the most successful techniques used employ message content initially, alongside other features.

3.2 User Profiles

Although it must be admitted that in creating a new Twitter account, personal information are limited, however, they can give beneficial insights for the SES of particular users. Users' profiles contain a different number of metadata such as the user's biography, followers, name, and location. The expectation is that a user's biography offers an important source of demographic data. However, Twitter users' biography is left empty for 48% of users, and others do not supply good-quality information (Culotta et al., 2016). (Preoţiuc-Pietro et al., 2015a) use the profile information of the account to capture users with self-disclosed occupations by annotating the user description field. (Lampos et al., 2016) use also profile description field of UK Twitter users to search for occupation mentions. In order to infer the user's socioeconomic status, most studies use description field and attempt to search for related information given by a particular user. These data are also useful in order to validate other SES features inferred from tweet messages.

3.3 Social Network Relations

The followers of a user represent a good indicator of their SES. Following reciprocal relationship can provide evidence of strong user connection. Some indicators can group regular exchanges of messages or frequent mention to names in messages. The number of tweets, mentions, links, hashtags and retweets, the number of followers, friends and the ratios of tweets to retweets are considered as statistical features. (Lampos et al., 2016) use these features to compile a set of latent topics that Twitter users were communicating. (Culotta et al., 2016) use the Twitter REST API and *followers/ids* request to sample many followers for each account, and the results are ordered with the most recent following first. And with the same methods, they use *friends/ids* API request to collect a list of friends. The example of (Barberá, 2016) best illustrates this idea it enables to overcome the collection of information about the entire network of a particular user that is costly and requiring multiple API calls, focuses on verified accounts. (Ikeda et al., 2013) use a community-based method with the extraction of the community from follower/followee relations followed by estimation of the demographics of the extracted communities. The demographic category of each community group is estimated using text-based method and the use of Fast Modularity Community Structure Inference Algorithm. Some studies assume that people within a given social class tend to have similar lifestyles using their income levels and common experience. Their interaction is called homophily. In the same context (Aletras and Chamberlain, 2018) use the information extracted from the extended networks of Twitter users in order to predict their occupational class and their income. They demonstrated that user's social network and their language use are complementary.

3.4 Spatial Information

The majority of smartphones are now equipped with Global Positioning System (GPS) functions and they work with geo-satellites which accurately infer the user's location with latitudes and longitudes coordinates. This would be an optional field for a particular user to enable due to their privacy choice. This indicator is very helpful when the person is mobile and usually updates their location profile. (Bokányi et al., 2017) obtained 63 million of Twitter geolocated messages from the area of the United States and assigned a county to each tweet. Once aggregated, daily tweeting activity allows to measure human activities and constitutes an important socioeconomic indicator whether a particular user is employed or not. In order to build a social class dataset, some studies attempt to show that the wealthier the place, the richer the users who usually visit it. (Mi-

randa Filho et al., 2014) used the lifestyle and the wealth of neighborhood people typically visit to label Brazilian users into various social classes. Then, they utilized Foursquare to label places according to the wealth of the neighborhood. They selected users who had at least one Foursquare interaction (Foursquare interactions include check-in (the user told a friend he/she was at a given place), tips (the user posts tips and opinion about a given place) and mayorship (title given to the most frequent user in a given location in the past 60 days). (Zhong et al., 2015) investigate the predictive power of location and the mobility to infer users' demographics with the use of location to profile (L2P) framework. The data crawling module accumulates user profiles and location check-in with corresponding information on Sine Weibo.

3.5 Temporal Information

Twitter enables researchers to analyze human activities during the 24 hours of the day because they are biologically bound to exhibit daily periodic behavior. In this context (Bokányi et al., 2017) aggregate monday to friday relative tweeting activities for each hour in each US County to form an average workday activity pattern, assuming that the activity patterns form a linear subspace of the 24-hour "time-space". This study shows that this measure correlates with county employment and unemployment rates in relation to lifestyles connected to regular working hours. The relationship between daily activity patterns and employment data can be captured using Twitter data.

3.6 Demographic Attributes

Some researchers attempted to include demographics as features. Age, for example, has a vital role in income prediction. Old people earn significantly more than young ones. Higher age leads to, on average, more work experience and education, which is translated into higher income. (Flekova et al., 2016) explored the relationship between stylistic and syntactic features, authors' age, and income, to conclude that the hypothesis of numerous feature type writing style and age use is predictive of income.

4 Inference Methods for SES Evaluation on Twitter

Different techniques have been used in the past and are being employed now to improve the accuracy of SES inference methodologies and algorithms. This burgeoning field lends techniques ranging from different areas of study involving machine learning, statistics, natural language processing to regression models. Various methods achieved different levels of success. The effectiveness and granularity levels produced by these methods continue to be improved.

Most recent researchers use a three-step methodology to infer the SES. First, they collect available information about a number of Twitter users. Secondly, they develop the classification method using additional data (number of followers, the content of tweets). And finally, they classify users who do not provide any concrete information according to SES. (Preoţiuc-Pietro et al., 2015a) for example, extracts occupation information from Twitter user profiles and uses text analysis to categorize users into occupational classes.

In general, a common approach to demographic inference is supervised classification, from a training set of labeled users, a model is fit to predict user features from the content of their writings. In other words, inferring user characteristics is framed as a predictive task validated on held-out data. This is done by establishing regression or classification methods.

4.1 Regression Methods

Various techniques for the inference of SES of Twitter users have been adopted from data mining and machine learning techniques. Some studies used the linear regression method, others used non-linear regression method and a third party used a hybrid approach that combines both linear and non-linear methods. A standard non-linear method does not inform which features are the most important in the predictive task. Then, the interpretability of linear methods allows performing an extensive qualitative and quantitative analysis of the input features. (Flekova et al., 2016) used both linear with Elastic Net regularization methods and non-linear with Support Vector regression together with an RBF kernel method. The authors found that machine learning regression methods can be used to predict and analyze user's income. (Lampos et al., 2016) used a non-linear generative learning approach, which consists of Gaussian Process (GP) and Kernel, to classify Twitter users according to SES as having upper, middle or lower level. Further, in (Preoţiuc-Pietro et al., 2015b), the authors used similar methods to study the user behavior and its power to predict income. It is important to note that GPs is a Bayesian non-

parametric statistical framework that formulates priority functions. (Hasanuzzaman et al., 2017) used linear and non-linear methods. The linear method is a logistic regression with Elastic Net regularization. In order to capture the non-linear relationship between a user's temporal orientation and their income, the authors used GP for regression. (Culotta et al., 2016) used a regression model for the prediction in order to understand the demographics of users. Due to the high dimensionality of features, the authors used elastic net regularization. Since each output variable consists of subject categories of demographic characteristics. They used a multitask variant of the elastic net to ensure the same features as selected for each category.

4.2 Classification Methods

(Mentink, 2016) employed two different approaches to classify the users in the dataset. The first is named the individual approach, it determines the performing classifier per feature group and consequently combines them via a soft-voting ensemble method. The second is named the combined approach, it calculates the performance scores for all possible combinations of classifiers and their respective ensemble (also via soft-voting). The author used Logistic Regression, Support Vector Machines, Naive Bayes and Random Forest algorithms. The author runs the algorithm to determine what occupation and what education-level label should be given to a particular user, to overcome the data imbalance, noise and bias, (Chen and Pei, 2014) used a typical imbalance classification approach which uses multiple classifier systems (MCS) and a sampling method which is a class-based random sampling method an extension of random under-sampling. The objective is to classify users according to their occupation. (Miranda Filho et al., 2014) evaluated a large number of classifiers using their WEKA version to generate classification models, including multinomial Naive Bayes (MNB), Support Vector Machine (SVM), and Random Forest. As MNB is more efficient than other algorithms, the authors used this method to infer social class for each particular user.

5 Tweet Gathering and Analysis

Messages on Twitter are publicly accessible in the online domain and can be gathered for study purposes. This availability makes Twitter an efficient tool in retrieving and analyzing public messages by allowing its users to become social sensors within the population.

5.1 Data Corpuses and Ressources

The corpus size of tweets grouped have varied from relatively small datasets to as large as three billion tweets (Mentink, 2016). The time span of the data collected was usually in the range of a few weeks to a couple of months, and sometimes a year. Table 1 shows some datasets and their sizes over the past years. First, the REST API is helpful for gathering particular user tweets, allowing the backtracking of their timeline. For example, to collect their most recent 3.200 tweets. Second, the streaming API that manages the tweets as they are being broadcast would only be able to receive 1% of the Firehose. Twitter data partners furnish a premium service that supplies messages covering a longer duration as well as 100% access to the Firehose. (Preoţiuc-Pietro et al., 2015a) created a publicly available data-set [4] of users, including their profile information and historical text content as well as a label to their occupational class from the "Standard Occupational Classification" taxonomy.

This public available dataset used by several researchers containing a group of $5,191$ users in total. However, the extraction of social network information of some accounts are not allowed. These accounts may have been annulled or become private. For example (Aletras and Chamberlain, 2018) reported results of $4,625$ users, from the original subset, that are still publicly available. Various studies (Preoţiuc-Pietro et al., 2015b; Lampos et al., 2016; Preoţiuc-Pietro et al., 2015a; Flekova et al., 2016) in the dataset creation mapped Twitter users to their income or their job title using standardized job classification taxonomy. The Standard Occupational Classification (SOC) is a UK-governmental system developed by the Office of National Statistics (ONS) for listing and grouping occupations. Jobs are organized hierarchically based on skill requirements and content.

(Culotta et al., 2016) mapped Twitter users according to their educational level (No College, College, Grad School) and other traits using Quantcast.com, an audience measurement society that tracks the demographics of users of millions of websites. The estimated demographics of a large number of sites are publicly accessible through the

[4] https://sites.sas.upenn.edu/danielpr/data

References	Corpus size	Period Covered	Corpus Origin
(Miranda Filho et al., 2014)	15.435 Users	Sep'13-Oct'13	Brazilian
(Preoţiuc-Pietro et al., 2015b)	10.796.836	Aug'14	US
(Barberá, 2016)	1.000.000.000	Jul'13-May'14	US
(Lampos et al., 2016)	2.082.651	Feb'14-Mar'15	US
(Mentink, 2016)	3.000.000.000	Nov'14-Oct'15	Dutch
(Hu et al., 2016)	9.800 Users		US
(Bokányi et al., 2017)	63.000.000	Jan'14 and Oct'14	US
(van Dalen et al., 2017)	2.700.000	Sep'16	Dutch
(Abitbol et al., 2018)	170.000.000	Jul'14-May'17	French
(Levy Abitbol et al., 2019)	90.369.215	Aug'14-Jul'15	French

Table 1: Datasets and Collection Periods of Some Studies.

use of searchable web interface. For each variable, Quantcast gives the expected percentage of visitors to a website with a given demographic.

5.2 Results and Metrics

The conclusions reached by different studies have been significantly improved over time with regards to increased accuracy and other measurements. Table 2 shows some techniques and their results over the past years. This has been driven by improvements in algorithms and inclusion of more useful features. It is important to note that the effectiveness and the reliability of occupation representativeness increase when estimating profession, using non-standard and out-of-vocabulary (OOV) occupation names. In this context, to overcome the limitation of the work of (Sloan et al., 2015) and (Mac Kim et al., 2016), (Kim et al., 2016) builded a machine learning model attempts to capture linguistically noisy or open-ended occupations in Twitter. This induces in more reliable occupation representativeness.

Different approaches have been introduced to compare the performance and results of the methods. They include accuracy and use of two other standard metrics: Pearson's correlation coefficient and Mean Absolute Error (MAE). To validate the effectiveness of the approaches against different baselines, the k-fold cross validation has been well utilized for precision, recall and F-measure: the standard metrics for classification methods.

Over time, accuracy levels of results have continued to be improved starting from 2013 when the inference of users' occupation was used. taking into consideration other information such as Twitter links, friends, user tweets, profiles, and other metadata associated with the message. Furthermore, with the adoption of various features such as user profile features, users psycho-demographic features, or user emotion features, accuracy has improved with the recent studies of (Mentink, 2016; Lampos et al., 2016) achieving a 75% accuracy.

6 Conclusion and Future Prospectives

The study of socioeconomic status inference is one of the most active field of information retrieval. Such works are positioned at a crossroads of multiple disciplines.

Different studies that introduced the inference of hidden user characteristics (Al Zamal et al., 2012; Miranda Filho et al., 2014; Volkova et al., 2015) are salient in the field. The results of these works are not only of interest to statistics agencies but also necessary for studies in the social science (targeted advertising, personalized recommendations of user posts and the possibility of extracting authoritative users (Pennacchiotti and Popescu, 2011)). It is important to introduce the role of SES in politics such as the works of (Barberá and Rivero, 2015), (Burckhardt et al., 2016), (Kalsnes et al., 2017), (Vargo and Hopp, 2017) and (Brown-Iannuzzi et al., 2017). Twitter is increasingly considered as politically transformative communication technology that allows citizens and politicians to connect, communicate, and interact easily (Chadwick, 2006). The flaw in previous studies of political behavior using Twitter data is the lack of information about the sociodemographic characteristics of individual users. Policy makers have recently suggested introducing well-being community which will help governments do a better job at directing public policy towards promoting quality of life.

The inference of SES is an ambitious problem as it may belong to a combination of environmen-

References	Technique	Accuracy (%)	Class
(Ikeda et al., 2013)	Hybrid Method	71.60	Occupation
(Siswanto and Khodra, 2013)	Machine Learning	77.00	Occupation
(Miranda Filho et al., 2014)	Machine Learning	73.00	Social Class
(Preoţiuc-Pietro et al., 2015a)	Gaussian Process	52.70	Occupation
(Mentink, 2016)	Hybrid Method	75.00	SES
(Lampos et al., 2016)	Gaussian Process	75.00	SES
(Poulston et al., 2016)	SVM Classifier	50.47	SES
(van Dalen et al., 2017)	Logistic Regression	72.00	Income
(Hasanuzzaman et al., 2017)	Supervise Learning	74.40	Income
(Aletras and Chamberlain, 2018)	Gaussian Process Regression (GPR)	50.44	Occupation

Table 2: Results and Techniques Used Over the Past Years.

tal variables and individual characteristics. Some of these characteristics can be easier determined like gender or age while others, are sometimes complicated with privacy issues and relying to some degree on self-definition, are harder to determine like occupation, ethnicity, education level or home location. Neverthless, there are many challenges in the inference of SES for Twitter users. Manual classification and data sampling are time-consuming, hard process and not scalable. Models are learned by referring to a datasets which were manually labeled using Amazon Mechanical Turk at a high monetary cost. Another issue is that people often misrepresent themselves on various online social platforms. This can lead to false data interpretations which as a result can affect the accuracy of the research. Automated detection tools are based on the supposition that users will introduce information on their demographic background through profile information or metadata. While it is not possible to expect that all users do this, those who did were a random group of the Twitter population, then we would not expect to discover conflicts in prevalence rates for sociodemographic characteristics (Sloan et al., 2015). Another problem is that Twitter data cannot represent all the populace as discussed previously. (Sloan, 2017) treated the issue of using human validation to find the accuracy of methods applying profile data to assign users to occupational groups and, he deduced that this process could provide misclassifications due to users reporting their hobbies and interests rather than their actual occupations (e.g, writer, artist). Another limit is that deriving income statistics from job labels is not a suitable method.

Given the findings presented above, the following are important issues to address in future Twitter socio-demographic inference studies. First, there is a need to look at the relationship between a user's actual demographic characteristics and how demographic categorization tools classify that user as a function of how profile information is presented and a virtual identity constructed. To conclude, there is a need to link Twitter profiles and survey data. Researchers can start theorizing better working machine learning models to improve accuracy and scalability. In addition, the methodologies used in different research projects can be coupled to increase efficiency. Another purpose for future research projects is to construct a less human effort, low computational cost and focus on the construction of a stronger evaluation framework.

References

Jacob Levy Abitbol, Márton Karsai, Jean-Philippe Magué, Jean-Pierre Chevrot, and Eric Fleury. 2018. Socioeconomic dependencies of linguistic patterns in twitter: A multivariate analysis. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, pages 1125–1134.

Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. 2012. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. *ICWSM* 270:2012.

Nikolaos Aletras and Benjamin Paul Chamberlain. 2018. Predicting twitter user socioeconomic attributes with network and language information. *arXiv preprint arXiv:1804.04095* .

Aaron Antonovsky. 1967. Social class, life expectancy and overall mortality. *The Milbank Memorial Fund Quarterly* 45(2):31–73.

Pablo Barberá. 2016. Less is more? how demographic sample weights can improve public opinion esti-

mates based on twitter data. Technical report, Working Paper.

Pablo Barberá and Gonzalo Rivero. 2015. Understanding the political representativeness of twitter users. *Social Science Computer Review* 33(6):712–729.

Eszter Bokányi, Zoltán Lábszki, and Gábor Vattay. 2017. Prediction of employment and unemployment rates from twitter daily rhythms in the us. *EPJ Data Science* 6(1):14.

Jazmin L Brown-Iannuzzi, Kristjen B Lundberg, and Stephanie McKee. 2017. The politics of socioeconomic status: how socioeconomic status may influence political attitudes and engagement. *Current opinion in psychology* 18:11–14.

Philipp Burckhardt, Raymond Duch, and Akitaka Matsuo. 2016. Tweet as a tool for election forecast: Uk 2015. general election as an example. *En: Third annual meefing of the Asian Polifical Methodology Society in Beijing* .

Andrew Chadwick. 2006. Internet politics: States, citizens, and new communication technologies. *New York, NY* .

J Crosby Chapman and Verner Martin Sims. 1925. The quantitative measurement of certain aspects of socio-economic status. *Journal of Educational Psychology* 16(6):380.

Ying Chen and Bei Pei. 2014. Weakly-supervised occupation detection for micro-blogging users. In *Natural Language Processing and Chinese Computing*, Springer, pages 299–310.

Charles D Cowan, Robert M Hauser, R Kominski, Henry M Levin, S Lucas, S Morgan, and C Chapman. 2012. Improving the measurement of socioeconomic status for the national assessment of educational progress: A theoretical foundation. *National Center for Education Statistics. Retrieved from http://files. eric. ed. gov/fulltext/ED542101. pdf* .

Aron Culotta, Nirmal Kumar Ravi, and Jennifer Cutler. 2016. Predicting twitter user demographics using distant supervision from website traffic data. *Journal of Artificial Intelligence Research* 55:389–408.

Lucie Flekova, Daniel Preoţiuc-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)*. volume 2, pages 313–319.

Wendell L French. 1959. Can a man's occupation be predicted? *Journal of Counseling Psychology* 6(2):95.

Mohammed Hasanuzzaman, Sabyasachi Kamila, Mandeep Kaur, Sriparna Saha, and Asif Ekbal. 2017. Temporal orientation of tweets for predicting income of users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. volume 2, pages 659–665.

Erika Hoff, Brett Laursen, Twila Tardif, et al. 2002. Socioeconomic status and parenting. *Handbook of parenting Volume 2: Biology and ecology of parenting* 8(2):231–252.

Tianran Hu, Haoyuan Xiao, Jiebo Luo, and Thuyvy Thi Nguyen. 2016. What the language you tweet says about your occupation. In *Tenth International AAAI Conference on Web and Social Media*.

Kazushi Ikeda, Gen Hattori, Chihiro Ono, Hideki Asoh, and Teruo Higashino. 2013. Twitter user profiling based on text and community mining for market analysis. *Knowledge-Based Systems* 51:35–47.

Rosie Jones, Ravi Kumar, Bo Pang, and Andrew Tomkins. 2007. I know what you did last summer: query logs and user privacy. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM, pages 909–914.

Bente Kalsnes, Anders Olof Larsson, and Gunn Sara Enli. 2017. The social media logic of political interaction: Exploring citizens and politicians relationship on facebook and twitter. *First Monday* 22(2).

Sunghwan Mac Kim, Stephen Wan, and Cécile Paris. 2016. Occupational representativeness in twitter. In *Proceedings of the 21st Australasian Document Computing Symposium*. ACM, pages 57–64.

Evelyn M Kitagawa and Philip M Hauser. 1973. Differential mortality in the united states: A study in socioeconomic epidemiology .

Nancy Krieger, David R Williams, and Nancy E Moss. 1997. Measuring social class in us public health research: concepts, methodologies, and guidelines. *Annual review of public health* 18(1):341–378.

Vasileios Lampos, Nikolaos Aletras, Jens K Geyti, Bin Zou, and Ingemar J Cox. 2016. Inferring the socioeconomic status of social media users based on behaviour and language. In *European Conference on Information Retrieval*. Springer, pages 689–695.

Kristina Lerman, Megha Arora, Luciano Gallegos, Ponnurangam Kumaraguru, and David Garcia. 2016. Emotions, demographics and sociability in twitter interactions. In *ICWSM*. pages 201–210.

Jacob Levy Abitbol, Eric Fleury, and Márton Karsai. 2019. Optimal proxy selection for socioeconomic status inference on twitter. *Complexity* 2019.

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)*. volume 1, pages 165–174.

Yong Li, Mengjiong Qian, Depeng Jin, Pan Hui, and Athanasios V Vasilakos. 2015. Revealing the efficiency of information diffusion in online social networks of microblog. *Information Sciences* 293:383–389.

Sunghwan Mac Kim, Stephen Wan, Cécile Paris, Jin Brian, and Bella Robinson. 2016. The effects of data collection methods in twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*. pages 86–91.

Michael G Marmot, Manolis Kogevinas, and Maryann A Elston. 1987. Social/economic status and disease. *Annual review of public health* 8(1):111–135.

Fons Mentink. 2016. *Machine driven predictions of the socio-economic status of Twitter users*. Master's thesis, University of Twente.

Renato Miranda Filho, Guilherme R Borges, Jussara M Almeida, and Gisele L Pappa. 2014. Inferring user social class in online social networks. In *SNAKDD*. pages 10–1.

Paul L Morgan, George Farkas, Marianne M Hillemeier, and Steven Maczuga. 2009. Risk factors for learning-related behavior problems at 24 months of age: Population-based estimates. *Journal of abnormal child psychology* 37(3):401.

J Michael Oakes and Peter H Rossi. 2003. The measurement of ses in health research: current practice and steps toward a new approach. *Social science & medicine* 56(4):769–784.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pages 430–438.

Adam Poulston, Mark Stevenson, and Kalina Bontcheva. 2016. User profiling with geo-located posts and demographic data. In *Proceedings of the First Workshop on NLP and Computational Social Science*. pages 43–48.

Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015a. 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)*. volume 1, pages 1754–1764.

Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015b. Studying user income through language, behaviour and affect in social media. *PloS one* 10(9):e0138717.

Daniele Quercia, Jonathan Ellis, Licia Capra, and Jon Crowcroft. 2012. Tracking gross community happiness from tweets. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. ACM, pages 965–968.

Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, pages 1524–1534.

Peter Schmidt and Robert P Strauss. 1975. The prediction of occupation using multiple logit models. *International Economic Review* pages 471–486.

Elisafina Siswanto and Masayu Leylia Khodra. 2013. Predicting latent attributes of twitter user by employing lexical features. In *Information Technology and Electrical Engineering (ICITEE), 2013 International Conference on*. IEEE, pages 176–180.

Luke Sloan. 2017. Who tweets in the united kingdom? profiling the twitter population using the british social attitudes survey 2015. *Social Media+ Society* 3(1):2056305117698981.

Luke Sloan, Jeffrey Morgan, Pete Burnap, and Matthew Williams. 2015. Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. *PloS one* 10(3):e0115545.

Aaron Smith and Joanna Brenner. 2012. Twitter use 2012. *Pew Internet & American Life Project* 4.

Brendan M Sullivan, Gopikrishna Karthikeyan, Zuli Liu, Wouter Lode Paul Massa, and Mahima Gupta. 2018. Socioeconomic group classification based on user features. US Patent App. 15/221,587.

AB Van Berkel-Van Schaik and B Tax. 1990. Towards a standard operationalisation of socioeconomic status for epidemiological and socio-medical research. *Rijswijk: ministerie van WVC* .

Reinder Gerard van Dalen, Léon Redmar Melein, and Barbara Plank. 2017. Profiling dutch authors on twitter: Discovering political preference and income level. *Computational Linguistics in the Netherlands Journal* 7:79–92.

Chris J Vargo and Toby Hopp. 2017. Socioeconomic status, social capital, and partisan polarity as predictors of political incivility on twitter: a congressional district-level analysis. *Social Science Computer Review* 35(1):10–32.

Svitlana Volkova and Yoram Bachrach. 2016. Inferring perceived demographics from user emotional tone and user-environment emotional contrast. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 1567–1578.

Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. 2015. Inferring latent user properties from texts published in social media. In *AAAI*. pages 4296–4297.

Karl R White. 1982. The relation between socioeconomic status and academic achievement. *Psychological bulletin* 91(3):461.

Erik Olin Wright and G Ritzer. 2003. Encyclopedia of social theory.

Yuan Zhong, Nicholas Jing Yuan, Wen Zhong, Fuzheng Zhang, and Xing Xie. 2015. You are where you go: Inferring demographic attributes from location check-ins. In *Proceedings of the eighth ACM international conference on web search and data mining*. ACM, pages 295–304.