Profiling a Country Using Twitter User data: the case for Turkey

Authors: Ali Hürriyetoğlu, Berra Karayel, Melih Can Yardı, Mehmet Said Baltalar, Çiğdem Erol, Nur Sude Kuzucu, Zeynep Dicle Gülhan, Rumeysa Hanife Kars, ...

Abstract

We apply and compare several user data collection methods for building a database of the users in Turkey. This database provides insights on what is happening in Turkey. The lack of such a database limits the capacity and quality of any study that uses Twitter data due to restrictions on Twitter content. The user information was retrieved from followers and keyword searches on streaming and search APIs.

“Literature Review Notes” dökümanına bak.

Introduction

In cases of intention to understand what is happening in a country or what do people of the country think about an issue using social media data, the relation between the population of the country and the subset of the population that generates this data must be known as much as possible. Which part of the population uses the platform? How is the reality reflected in the posts? Are there significant differences among users on the platform? How does the frequency of the posts from a user affect results? Is it possible to draw a sample that is representative of the actual population in the country from the platform users? Answers to these questions ensure the quality of any analysis using data collected from the platform. Therefore, we study the quality of various user information collection methods from Twitter and analyze their utility for determining what is happening in Turkey.

We investigate various methods such as retrieving followers of users, filtering tweets collected from streaming and search APIs to collect user data from Twitter and create a database of users in Turkey. The user database is analyzed for various characteristics and compared with external knowledge related to the population in Turkey. Moreover, same analyses were performed for each user subset collected using one of the aforementioned different methods. Finally, methods for sampling were analyzed to determine the one that yields the most representative sample.

Relevant work

Collecting tweets in the Dutch language in order to measure the characteristics of the Dutch population was achieved by Tjong Kim Sang and van den Bosch (2013). The authors utilize a stopword list in Dutch to identify tweets in Dutch.

Data Collection

We apply three methods for collecting user profiles from Twitter. These methods are follower information, user objects from the posts collected from the Twitter Streaming API, and user objects from the posts collected using Twitter Search API.

We evaluate the quality of the sample by seeking answers to the following questions:
1- Do subsets retrieved using different methods overlap?
2- Does the information extracted from users correlate with information in the official information sources? Some of the possibilities are distribution of the age, gender, occupation, education level of the users across districts in Turkey.
3- Are there surveys that can be replicated using this data in raw or stratified form?
4- What is an ideal size for a random sample? We can compare samples with the same and various sizes in terms of the characteristics of the user they contain.

Use cases

1- Profiling Political Behavior of Turkey
2- Twitter User Types
3- Location Identification
4- How can we Identify Gender on Twitter Data: Turkey Example
5- Twitter Age Identification
6- Interpretive Quantitative Sociology

1- Profiling Political Behavior of Turkey

Country profiles enable us to understand a country, its people, cities, villages, culture, and varied characteristics. The goal of country profiling is to look at it from inside out. The approach is looking from outside to understand a country better.[1] While we are profiling a characteristic of a country, we need to identify some certain indicators. In this research, we need to search for some key indicators, broad categories which affect the political tendency in Turkey.[2] There are several indicators that are used in the similar researches such as political

[1] https://guides.nyu.edu › CountryData › profiles-reports
[2] http://www3.weforum.org/docs/GGGR12/UsersGuide_GGGR12.pdf

stability and violence, permanence of regime type, refugees produced, employment, unemployment, expenditure, taxes, government deficits and debts and so on.[3] In addition to that for the political indicators, Princeton University created five main groups which are country background, democracy and regime type, parties, parliaments, governments, and leaders, governance indicators, regional/subnational governments & federalism.[4] At that point, we need to decide our indicators based on our sample and then look for some international indexes or political participation indexes which are represented by public international sources and give the best statistics available for those indicators at the time of the preparation of the human development report (UNDP). There are some country statistical profiles in OECD, BBC News, Yale University Library, UNCTAD (United Nations Conference on Trade and Development) and so on. They provide guidance to the political background of countries. They might be useful for us regarding the indicators that we will determine.

Most researches in the literature examines individual-level participation, comparing the activity of different groups according to variables such as age, sex, occupation, education, ethnicity, and family background. One study examines levels of participation across a wide variety of different types – political and social, conventional and unconventional, formal and informal, individual and collective (Newton & Giebler, 2008). The study is divided into two parts, in the first part, they show the patterns of individual-level participation, and the second part explores the theories that might account for these patterns of participation. Researchers suggest that each country's political history is associated with its individualist or collectivist based on culture. Besides, others emphasize the importance of religion, urban-rural differences, government structures, contextual factors of opportunity structures and modernization, and of age, education, and wealth (Newton & Giebler, 2008) which might be very useful for us in this research. We can combine those categories with Mustafa Aydın's work[5] that focuses on social and political tendencies of Turkey which I will explain in the next paragraph.

One of the most detailed work on political trends in Turkey is the one which was done by Mustafa Aydın, Kadir Has University. This is a longitudinal study so that we can access different time periods and compare them with each other. They conduct face to face interviews (age 18 and over, residing in city centers of 26 cities, questions on politics, economic developments, social relations, and international issues.). January 2021 Quantitative Research Report states the political, religious and ethnic background of the respondents after giving some demographics. In this report, we see the positions of respondents in the political spectrum. The question asked is "How would you define your political views?"These figures include the years between 2015-2019. It also gives the distribution of age in this political spectrum. In addition to that they asked: "how would you position your views in political matters?" and showed a figure. Then, in this report we can see the distribution of left-right spectrum by age, socioeconomic status, and by political parties. There are some figures for nationalism. The question for that: "In general, to which extent do you define your nationalism level?". According to the previous report (Jan, 2020), the rate of those who think that there is political polarization in Turkey has reached

[3] https://carleton.ca/cifp/governance-democracy-processes/indicator-descriptions/
[4] https://libguides.princeton.edu/politics/indicators#s-lg-box-11530398
[5] http://www.mustafaaydin.gen.tr/20/turkiye-sosyal-siyasal-egilimler-political-and-social-trends-in-turkey

the highest rank of the last 4 years with 55.9%. As the axis of polarization, the secular-religious divide stands out with 42.9% and the right-left divide with 28.6%. On the other hand, the differences in the view of government policies, the country's preferred political system and support for democracy also point to political polarization.

Concerning political behavior of Turkish people, voting behavior should also be observed. In order to understand voting behavior, I searched for some research companies like Konda, Sonar Araştırma, Asarda and SosyoPolitik Saha Araştırmaları Merkezi. These studies shows turkey-wide political agenda and tendency of voters in Turkey by using surveys. They asked for the current system, which party they would choose to vote for, which politician they support, and so on. Also, Tüik gives some demographic information about the voters in Turkey and the results of the previous elections (further info from TÜİK will be elaborated.)

2- Twitter User Types[6]

The tweets of different twitter user types have some characteristic features. Studies analyzing Twitter API data divided users into various categories and used different methods to make this distinction. In research conducted using Twitter user data, it is important to identify and distinguish appropriate categories to make inferences. Various methods like textual content analyses or tweeting behavior can be used to distinguish between user types. The studies carried out so far provide us valuable information about twitter user behavior.

In a research conducted by Chu et al. (2012) twitter users are divided into three categories as human, bot, and cyborg. The main aim of their work is to distinguish real users from bots. Human assisted bots and bot assisted human users (cyborgs) are not as easy to distinguish as other types as they possess interweave characteristics of both manual and automated behavior. To develop an automated classification system, the study creates a ground-truth set that contains known samples of human, bot, and cyborg and classifies them by manually checking their user logs and homepages. Tweet contents, visit URLs included in tweets, tweeting devices, user profile, and the numbers of followers and friends are considered as indicators of user type behavior. In textual analyzes it is seen that humans tend to produce original, intelligent content rather contents that lack originality and complexity, the attributes we encounter in bots. Also, study shows that more than half of the human tweets are manually posted via the Twitter website. In general, tweeting via devices such as Tweetie, UberTwitter, Mobile web… etc. requires human participation. In contrast, the top tools used by bots are mainly auto piloted, and 42.39 percent of bot tweets are generated via unregistered API-based tools. It is shown that spammers give more external URLs, have fewer followers than those they follow, and send more messages than real users. Another study (Yardi et al. 2010) also confirms: According to their observations, spammers send more messages than legitimate users, and are more likely to follow other spammers than legitimate users. Thus, a high follower to following ratio is a sign of spamming behavior. In the study, account reputation calculation is

[6] https://github.com/cigdemerolll/SCQ-User-Type

done to distinguish real and spam users. Account reputation is found by dividing the number of followers by the sum of the followers and friend numbers. The account reputation of real accounts clusters around 0.5, while bot accounts concentrates around 0.3. [1] [2] Tweeting device and the sequence are other important variables to consider. It is observed that human users tweet with large interarrivals (such as hours, and even days for some inactive users), and manual behavior cannot generate tweeting frequency as high as a bot. Many bots are driven by timers to post tweets at fixed interarrivals, and thus exhibit regular behavior. In contrast, human behavior carries the inherent complexity.

In another research (Uddin et al. 2014) Twitter users are classified into different classes according to their profile and tweeting behavior information. Based on a manual investigation of randomly selected Twitter profiles, six broad classes of Twitter users are identified (personal, professional, business and spam users, feed/news, viral/marketing services). After the manual analysis phase, the information gathered on user types is used to train a machine learning classifier to automatically classify Twitter profiles. To annotate the collected dataset, a simple computer program (using Java language) is implemented to measure important profile specific statistics. Next, based on Profile and Tweeting Behavior Specific Information the profiles are manually classified into the six classes. For cases where some user qualifies to be categorized in more than two classes, their identity is cross checked by visiting URL provided in the description of their profiles. The results of this study show that individual accounts usually send their tweets from mobile devices and use an informal language in their tweets. Individual accounts use more emojis than other account types and have a tendency to express both positive and negative emotions in their tweets. They are interactive. It is seen that professional accounts are also interactive like personal accounts, and they use a lot of mentions. Also, follower and friend counts tend to be close to each other.[3] [4] Spam users, feed/news and marketing services accounts are considered as digital actors of twitter as they are controlled by automated bots. Common features to detect these types include highly frequent tweeting, no or less interactivity, and either increase or decrease of their followers over time.

The last work we will talk about is the work of Lalindra De Silva and Ellen Riloff (de Silva & Riloff, 2014). The aim of the study is to find an answer to the question of can the user type influence the event relevance of a tweet. This study divides twitter users into organizations and individuals. Textual content is used to determine the user type. A ‘person heuristic’ is created to tag a tweet as a person-tweet. In order to distinguish between personal and organizational accounts, the information written in twitter bios is taken into account. While organizations introduce themselves with words such as "agency","institute","company", individuals use their names and surnames (for this, English and Spanish name and surname data, which are the languages in which the study was conducted, were used), introduce themselves such as 'I am...'. A manual annotation task was performed to ensure the accuracy of person heuristic. In this study, as in other studies, it is seen that real individual users mostly tweet from a mobile device. In the linguistic features analysis, the presence of expressions containing emotion, swearing, and exclamation was examined (the presence of expressing emotions is a sign of individual users). Apart from that, the study shows that the user mention attribute and the frequency of hashtag usage are related to individual user behavior.

Related work on user type detection and classification includes different features to classify users and there are various dimensions in which users are classified. Some studies use linguistic, profile, and social network features to classify users into political affiliations, while some others use “list” feature of twitter to classify elite users as celebrities, bloggers, and representatives of media outlets and other formal organizations. Twitter user types should be categorized in different ways according to the focus of the study. The research done in this area produced valuable information regarding certain types of twitter users. In future work, it will be appropriate to make functional categorizations and to benefit from the methods and research results of existing studies in the literature.

3- Location Identification[7]

Introduction

Since Twitter is a platform that uses location information, this feature can enable us to draw important conclusions while examining the data. Tweets or user profiles that pass through the filtering process can rapidly bring us the results we are looking for on a map, district or specific location. Twitter customers frequently design products that need to use the location of a Tweet or the user who posted it. For example, a client may be interested in public opinion regarding healthcare legislation in a specific region of the country, or may want to monitor customer satisfaction in different regions. Alternatively, they may wish to investigate social media interactions during extreme weather events. Customers who want to use or integrate location data into their products encounter difficulties in determining the best type of data to employ. The level of precision and accuracy provided for different types of data, as well as the convenience of use in filtering for different types of data, are factors in this determination.

The geographic metadata comes with a Tweet, gives by the option to "geotag" a Tweet when posting it by users. This geotagging can be based on a specific location, an assigned Twitter Place, or both. Twitter Places can be thought of at the neighborhood level, providing a "bounding box" with latitude and longitude coordinates that define the location region. The maximum level of precision is provided by this sort of geographic metadata, known as "Tweet Location." To obtain geographic information, Tweet Locations do not require language parsing or processing. The biggest drawback of relying on Tweet Locations is that only 1-2% of Tweets are geotagged. Furthermore, targeting very large areas (e.g. the entire State or Province) necessitates the usage of an important set of PowerTrack rules. The place_country: Operator, on the other hand, makes it simple to filter for certain countries. Also, places offer nice options, including an option to filter by country code or place name.

A second source of geospatial metadata is the mention of locations in Tweet content. Parsing the Tweet message for location names of interest, including nicknames, is required for such 'Place Mentioned' metadata. One Tweet may mention Manhattan, while another may mention the Big Apple. These types of Tweets are fairly easy to use, provides you the information of how people on Twitter refer to where you interested in. You can apply keywords

[7] https://github.com/msbaltalar/LocationAnalysisSCQ.git

or phrases to search for these phrases. On the other hand, because it is a less trustworthy indicator of the user's precise location, accuracy is likely to be lower.

Lastly, every Twitter profile contains a “Location” setting that the account holder can fill out. These Profile Locations are the most comprehensive repository of geospatial metadata. This information is not available from everyone and can contain any expression the user wishes. The location of one Twitter account might be "Living in the Colorado outskirts," while another might be set to "My family's basement," which is less helpful. This form of reference is a midpoint - it's not a precise GPS-verified geographic spot, but it is determined by the user, giving an extra boost to the reliability expectation. There are several filtering possibilities for this type of data, which are detailed below.

In summary, geo-referenced tweets have three metadata sources. The first one is, Tweet Location that are the tweets geotagged with an exact location or a Twitter Place. An example for an exact location with long/latitude coordinates can be -85.7629, 38.2267 while a Twitter Place with four sets of latitude/longitude coordinates that define an island can be ("Louisville Central") and a "bounding box". The second is Mentioned Location which is Geolocation that determined by parsing the tweet message, such as “If you're in Louisville, check out the pizzeria off the main road” or "I'm in Louisville and it's raining cats and dogs". The third one is Profile Location which contains parsing account level location for places of interest, for instance “I live in Louisville, home of Derby!” or “I live in Louisville, in beautiful Colorado.”

The usage of this metadata to georeference Tweets are contains many ways to filter such geospatial metadata which Twitter PowerTrack[1] [2] provides. These filters or rules were created using more than fifty PowerTrack Operators. For an introduction to PowerTrack Operators that can be used to filter Tweet and Profile Locations, see our article “Filtering Twitter by location”. Because Profile Locations are by far the largest source of Twitter geo metadata, Twitter provides Profile Geo enrichment. Profile Geo substantially increases the amount of geospatial data, hence this enrichment has been widely utilized.

The results obtained after the location analysis within the scope of the SCQ project

In our project, after 'text mining' operations with the data we have, we obtained 16084 data that we can use over 70112 lines. We took this obtained data into ASCII format and evaluated it. As a result of these evaluations, we reached 3696 unique location information out of 16084 filled data and sorted it. According to the results we obtained; We obtained two different clusters of Turkey's cities and the data we used. Based on the figures we obtained, we determined the proportions of Turkey's population distribution. By looking at these ratios; We have seen information that some of the cities added by users are repeating at much different rates than would normally be expected. In fact, our current data has been adjusted and formulated to give us an analysis of any location we want to focus on. Being able to analyze

separately for a province, district and a specific location can provide us with new findings at many different points.

Related Sources Citation & Abstract

1. Foucaud, Florent & Mertzios, George & Naserasr, Reza & Parreau, Aline & Valicov, Petru. (2014). Identification, location-domination and metric dimension on interval and permutation graphs. I. Bounds. Theoretical Computer Science. 668. 10.1016/j.tcs.2017.01.006.

We consider the problems of finding optimal identifying codes, (open) locating–dominating sets and resolving sets of an interval or a permutation graph. In these problems, one asks to find a subset of vertices, normally called a *solution* set, using which all vertices of the graph are distinguished. The identification can be done by considering the neighborhood within the solution set, or by employing the distances to the solution vertices. Normally the goal is to minimize the size of the solution set then. Here we study the case of interval graphs, unit interval graphs, (bipartite) permutation graphs and cographs. For these classes of graphs we give tight lower bounds for the size of such solution sets depending on the order of the input graph. While such lower bounds for the general class of graphs are in logarithmic order, the improved bounds in these special classes are of the order of either quadratic root or linear in terms of number of vertices. Moreover, the results for cographs lead to linear-time algorithms to solve the considered problems on inputs that are cographs.

2. Chappell, P., Tse, M., Zhang, M., & Moore, S. (2017). Using GPS geo-tagged social media data and geodemographics to investigate social differences: A twitter pilot study. *Sociological Research Online, 22*(3), 38-56. https://doi.org/10.1177/1360780417724065

This article outlines a new method for investigating social position through geo-tagged Twitter data, specifically through the application of the geodemographic classification system Mosaic. The method involves the identification of a given tweeter's likely location of residence from the 'geo-tag' attached to their tweet. Using this high-resolution geographic information, each individual tweet is then attributed a geodemographic classification. This article shows that the specific application of geodemographics for discerning between different types of tweeters is problematic in some ways, but that the general process of classifying tweeters according to their position in geographical space is viable and represents a powerful new method for discerning the social position of tweeters. Further research is required in this area, as there is great potential in employing the mobile global positioning system data appended to digital by-product data to explore the intersections between geographical space and social position.

3. Burrows, R., Webber, R., & Atkinson, R. (2017). Welcome to ‘Pikettyville’? Mapping London’s alpha territories. The Sociological Review, 65(2), 184–201. https://doi.org/10.1111/1467-954X.12375

This paper considers the influence of the burgeoning global ‘super-rich’ on contemporary socio-spatialization processes in London in the light of a contemporary re-reading of Pahl’s classic volume, *Whose City?* It explores if a turn to ‘big data’ – in the form of commercial geodemographic classifications – can offer any additional insights to a sociological approach to the study of the ‘super-rich’ that extends the ‘spatialization of class’ thesis further ‘up’ the class structure.

4. Graham, M., Hale, S. A., & Gaffney, D. (2014). Where in the world are you? Geolocation and language identification in Twitter. *The Professional Geographer*, *66*(4), 568-578.

The movements of ideas and content between locations and languages are unquestionably crucial concerns to researchers of the information age, and Twitter has emerged as a central, global platform on which hundreds of millions of people share knowledge and information. A variety of research has attempted to harvest locational and linguistic metadata from tweets to understand important questions related to the 300 million tweets that flow through the platform each day. Much of this work is carried out with only limited understandings of how best to work with the spatial and linguistic contexts in which the information was produced, however. Furthermore, standard, well-accepted practices have yet to emerge. As such, this article studies the reliability of key methods used to determine language and location of content in Twitter. It compares three automated language identification packages to Twitter's user interface language setting and to a human coding of languages to identify common sources of disagreement. The article also demonstrates that in many cases user-entered profile locations differ from the physical locations from which users are actually tweeting. As such, these open-ended, user-generated profile locations cannot be used as useful proxies for the physical locations from which information is published to Twitter.

5. Graham, S. D. (2005). Software-sorted geographies. *Progress in human geography*, *29*(5), 562-580.

This paper explores the central role of computerized code in shaping the social and geographical politics of inequality in advanced societies. The central argument is that, while such processes are necessarily multifaceted, multiscaled, complex and ambivalent, a great

variety of `software-sorting' techniques is now being widely applied in efforts to try to separate privileged and marginalized groups and places across a wide range of sectors and domains. This paper's central demonstration is that the overwhelming bulk of software-sorting applications is closely associated with broader transformations from Keynesian to neoliberal service regimes. To illustrate such processes of software-sorting, the paper analyses recent research addressing three examples of software-sorting in practice. These address physical and electronic mobility systems, online geographical information systems (GIS), and face-recognition closed circuit television (CCTV) systems covering city streets. The paper finishes by identifying theoretical, research and policy implications of the diffusion of software-sorted geographies within which computerized code continually orchestrates inequalities through technological systems embedded within urban environments.

6. Harris, R., Sleight, P., & Webber, R. (2005). *Geodemographics, GIS and neighbourhood targeting* (Vol. 8). John Wiley & Sons.

Geodemographic classification is 'big business' in the marketing and service sector industries, and in public policy there has also been a resurgence of interest in neighbourhood initiatives and targeting. As an increasing number of professionals realise the potential of geographic analysis for their business or organisation, there exists a timely gap in the market for a focussed book on geodemographics and GIS.

7. Parker, S., Uprichard, E., & Burrows, R. (2007). Class places and place classes geodemographics and the spatialization of class. *Information, Communication & Society*, *10*(6), 902-921.

This paper argues that the ‘spatial turn’ in the sociology of class – the clustering of people with a similar habitus into what we might think of as ‘class places’ – is connected in a number of important ways with the ongoing informatization of place, particularly as manifest in the urban informatics technology of geodemographics. This is a technology concerned with the development of the classification of places to commercial and policy ends – the assigning of postcodes to a set of mutually exclusive and exhaustive categories, or ‘place classes’. What interests the authors is the manner in which there is a strong concordance between the conclusions of academic sociologists working on the spatialization of class and those of – what might be thought of as – ‘commercial sociologists’ working in the geodemographics industry. Although the conceptual argot is very different, both have in common an interest in the codification and spatial mapping of habitus, and both arrive at very similar substantive conclusions about contemporary processes of sociocultural spatial clustering. But the authors' interest is not just in the observation that there is an analytic convergence in academic and commercial concerns with the relationship between ‘class places’ and ‘place classes’; rather, it is in their possible co-construction. They argue that geodemographic classifications are not only sociologically important phenomena but also represent an interesting example of a new form of software-mediated recursive urban ontology.

8. Sloan, L., & Morgan, J. (2015). Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. *PloS one*, *10*(11), e0142209.

In this paper we take advantage of recent developments in identifying the demographic characteristics of Twitter users to explore the demographic differences between those who do and do not enable location services and those who do and do not geotag their tweets. We discuss the collation and processing of two datasets—one focusing on enabling geoservices and the other on tweet geotagging. We then investigate how opting in to either of these behaviours is associated with gender, age, class, the language in which tweets are written and the language in which users interact with the Twitter user interface. We find statistically significant differences for both behaviours for all demographic characteristics, although the magnitude of association differs substantially by factor. We conclude that there are significant demographic variations between those who opt in to geoservices and those who geotag their tweets. Not withstanding the limitations of the data, we suggest that Twitter users who publish geographical information are not representative of the wider Twitter population.

9. Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P., & Rana, O. (2013). Knowing the Tweeters: Deriving sociologically relevant demographics from Twitter. *Sociological research online*, *18*(3), 74-84.

A perennial criticism regarding the use of social media in social science research is the lack of demographic information associated with naturally occurring mediated data such as that produced by Twitter. However the fact that demographics information is not explicit does not mean that it is not implicitly present. Utilising the Cardiff Online Social Media ObServatory (COSMOS) this paper suggests various techniques for establishing or estimating demographic data from a sample of more than 113 million Twitter users collected during July 2012. We discuss in detail the methods that can be used for identifying gender and language and illustrate that the proportion of males and females using Twitter in the UK reflects the gender balance observed in the 2011 Census. We also expand on the three types of geographical information that can be derived from Tweets either directly or by proxy and how spatial information can be used to link social media with official curated data. Whilst we make no grand claims about the representative nature of Twitter users in relation to the wider UK population, the derivation of demographic data demonstrates the potential of new social media (NSM) for the social sciences. We consider this paper a clarion call and hope that other researchers test the methods we suggest and develop them further.

10. Uprichard, E., Burrows, R., & Parker, S. (2009). Geodemographic code and the production of space. *Environment and Planning A*, *41*(12), 2823-2835.

There is a growing body of research relating to the ways in which digital code contributes to the production of space. In much of this work this issue is approached by first examining particular spaces and then considering the code and its effects on those spaces. In contrast, we explore the production of space from another angle, examining the ways in which an example of code—geodemographic classification—is constructed, and then questioning what it is about the emergent production of space that may feed back recursively into the production of that code.

11. Webber, R., & Butler, T. (2007). Classifying pupils by where they live: how well does this predict variations in their GCSE results?. *Urban Studies*, *44*(7), 1229-1253.

This paper summarises key findings resulting from the appending of the neighbourhood classification system Mosaic to the records of the Pupil Level Annual School Census (PLASC) within the National Pupil Database (NPD) of the Department for Education and Skills (DfES). The most significant of these findings is that, other than the performance of the pupil at an earlier Key Stage test, the type of neighbourhood in which a pupil lives is a more reliable predictor of a pupil's GCSE performance than any other information held about that pupil on the PLASC database. Analysis then shows the extent to which the performance of pupils from any particular type of neighbourhood is also incrementally affected by the neighbourhoods from which the other pupils in the school they attend are drawn. It finds that whilst a pupil's exam performance is affected primarily by the social background of people he or she may encounter at home, the social background of fellow school pupils is of only marginally lower significance. These findings suggest that so long as pupils' GCSE performances are so strongly affected by the type of neighbourhood in which they live, a school's league position bears only indirect relationship to the quality of school management and teaching. A better measurement of the latter would be a league table system which took into account the geodemographic profile of each school's pupil intake. The paper concludes with discussion of the relevance of these findings to the sociology of education, to the debate on consumer choice in public services, to the general appropriateness of adjusting public-sector performance metrics to take into account the social mix of service users and to parental strategies in the educational sector in particular.

12. Mourad A., Scholer F., Sanderson M. (2017) Language Influences on Tweeter Geolocation. In: Jose J. et al. (eds) Advances in Information Retrieval. ECIR 2017. Lecture Notes in Computer Science, vol 10193. Springer, Cham. https://doi.org/10.1007/978-3-319-56608-5_26

We investigate the influence of language on the accuracy of geolocating Twitter users. Our analysis, using a large corpus of tweets written in thirteen languages, provides a new understanding of the reasons behind reported performance disparities between languages. The results show that data imbalance has a greater impact on accuracy than geographical coverage.

A comparison between *micro* and *macro* averaging demonstrates that existing evaluation approaches are less appropriate than previously thought. Our results suggest both averaging approaches should be used to effectively evaluate geolocation.

13. Sakaki, T., Okazaki, M., & Matsuo, Y. (2010, April). Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web* (pp. 851-860).

Twitter, a popular microblogging service, has received much attention recently. An important characteristic of Twitter is its real-time nature. For example, when an earthquake occurs, people make many Twitter posts (tweets) related to the earthquake, which enables detection of earthquake occurrence promptly, simply by observing the tweets. As described in this paper, we investigate the real-time interaction of events such as earthquakes in Twitter and propose an algorithm to monitor tweets and to detect a target event. To detect a target event, we devise a classifier of tweets based on features such as the keywords in a tweet, the number of words, and their context. Subsequently, we produce a probabilistic spatiotemporal model for the target event that can find the center and the trajectory of the event location. We consider each Twitter user as a sensor and apply Kalman filtering and particle filtering, which are widely used for location estimation in ubiquitous/pervasive computing. The particle filter works better than other comparable methods for estimating the centers of earthquakes and the trajectories of typhoons. As an application, we construct an earthquake reporting system in Japan. Because of the numerous earthquakes and the large number of Twitter users throughout the country, we can detect an earthquake with high probability (96% of earthquakes of Japan Meteorological Agency (JMA) seismic intensity scale 3 or more are detected) merely by monitoring tweets. Our system detects earthquakes promptly and sends e-mails to registered users. Notification is delivered much faster than the announcements that are broadcast by the JMA.

14. Sadilek, A., Kautz, H., & Bigham, J. P. (2012, February). Finding your friends and following them to where you are. In *Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 723-732).

Location plays an essential role in our lives, bridging our online and offline worlds. This paper explores the interplay between people's location, interactions, and their social ties within a large real-world dataset. We present and evaluate Flap, a system that solves two intimately related tasks: link and location prediction in online social networks. For link prediction, Flap infers social ties by considering patterns in friendship formation, the content of people's messages, and user location. We show that while each component is a weak predictor of friendship alone, combining them results in a strong model, accurately identifying the majority of friendships. For location prediction, Flap implements a scalable probabilistic model of human mobility, where we treat users with known GPS positions as noisy sensors of the location of their friends. We explore supervised and unsupervised learning scenarios, and focus on the efficiency of both learning and inference. We evaluate Flap on a large sample of highly active users from two distinct geographical areas and show that it (1) reconstructs the entire friendship graph with high accuracy even when no edges are given; and (2) infers people's fine-grained location, even

when they keep their data private and we can only access the location of their friends. Our models significantly outperform current comparable approaches to either task.

15. Roller, S., Speriosu, M., Rallapalli, S., Wing, B., & Baldridge, J. (2012, July). Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 1500-1510).

The geographical properties of words have recently begun to be exploited for geolocating documents based solely on their text, often in the context of social media and online content. One common approach for geolocating texts is rooted in information retrieval. Given training documents labeled with latitude/longitude coordinates, a grid is overlaid on the Earth and pseudo-documents constructed by concatenating the documents within a given grid cell; then a location for a test document is chosen based on the most similar pseudo-document. Uniform grids are normally used, but they are sensitive to the dispersion of documents over the earth. We define an alternative grid construction using k-d trees that more robustly adapts to data, especially with larger training sets. We also provide a better way of choosing the locations for pseudo-documents. We evaluate these strategies on existing Wikipedia and Twitter corpora, as well as a new, larger Twitter corpus. The adaptive grid achieves competitive results with a uniform grid on small training sets and outperforms it on the large Twitter corpus. The two grid constructions can also be combined to produce consistently strong results across all training sets.

16. Priedhorsky, R., Culotta, A., & Del Valle, S. Y. (2014, February). Inferring the origin locations of tweets with quantitative confidence. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 1523-1536).

Social Internet content plays an increasingly critical role in many domains, including public health, disaster management, and politics. However, its utility is limited by missing geographic information; for example, fewer than 1.6% of Twitter messages (tweets) contain a geotag. We propose a scalable, content-based approach to estimate the location of tweets using a novel yet simple variant of gaussian mixture models. Further, because real-world applications depend on quantified uncertainty for such estimates, we propose novel metrics of accuracy, precision, and calibration, and we evaluate our approach accordingly. Experiments on 13 million global, comprehensively multi-lingual tweets show that our approach yields reliable, well-calibrated results competitive with previous computationally intensive methods. We also show that a relatively small number of training data are required for good estimates (roughly 30,000 tweets) and models are quite time-invariant (effective on tweets many weeks newer than the training set). Finally, we show that toponyms and languages with small geographic footprint provide the most useful location signals.

17. Kinsella, S., Murdock, V., & O'Hare, N. (2011, October). " I'm eating a sandwich in Glasgow" modeling locations with tweets. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents* (pp. 61-68).

Social media such as Twitter generate large quantities of data about what a person is thinking and doing in a particular location. We leverage this data to build models of locations to improve our understanding of a user's geographic context. Understanding the user's geographic context can in turn enable a variety of services that allow us to present information, recommend businesses and services, and place advertisements that are relevant at a hyper-local level.

In this paper we create language models of locations using coordinates extracted from geotagged Twitter data. We model locations at varying levels of granularity, from the zip code to the country level. We measure the accuracy of these models by the degree to which we can predict the location of an individual tweet, and further by the accuracy with which we can predict the location of a user. We find that we can meet the performance of the industry standard tool for predicting both the tweet and the user at the country, state and city levels, and far exceed its performance at the hyper-local level, achieving a three- to ten-fold increase in accuracy at the zip code level.

18. Jurgens, D., Finethy, T., McCorriston, J., Xu, Y. T., & Ruths, D. (2015, April). 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*.

Geolocated social media data provides a powerful source of information about place and regional human behavior. Because little social media data is geolocation-annotated, inference techniques serve an essential role for increasing the volume of annotated data. One major class of inference approaches has relied on the social network of Twitter, where the locations of a user's friends serve as evidence for that user's location. While many such inference techniques have been recently proposed, we actually know little about their relative performance, with the amount of ground truth data varying between 5% and 100% of the network, the size of the social network varying by four orders of magnitude, and little standardization in evaluation metrics. We conduct a systematic comparative analysis of nine state-of-the-art network-based methods for performing geolocation inference at the global scale, controlling for the source of ground truth data, dataset size, and temporal recency in test data. Furthermore, we identify a comprehensive set of evaluation metrics that clarify performance differences. Our analysis identifies a large performance disparity between that reported in the literature and that seen in real-world conditions. To aid reproducibility and future comparison, all implementations have been released in an open source geoinference package.

19. Cheng, Z., Caverlee, J., & Lee, K. (2010, October). You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 759-768).

We propose and evaluate a probabilistic framework for estimating a Twitter user's city-level location based purely on the content of the user's tweets, even in the absence of any other geospatial cues. By augmenting the massive human-powered sensing capabilities of Twitter and related microblogging services with content-derived location information, this framework can overcome the sparsity of geoenabled features in these services and enable new locationbased personalized information services, the targeting of regional advertisements, and so on. Three of the key features of the proposed approach are: (i) its reliance purely on tweet content, meaning no need for user IP information, private login information, or external knowledge bases; (ii) a classification component for automatically identifying words in tweets with a strong local geo-scope; and (iii) a lattice-based neighborhood smoothing model for refining a user's location estimate. The system estimates k possible locations for each user in descending order of confidence. On average we find that the location estimates converge quickly (needing just 100s of tweets), placing 51% of Twitter users within 100 miles of their actual location.

20. Backstrom, L., Sun, E., & Marlow, C. (2010, April). Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web* (pp. 61-70).

Geography and social relationships are inextricably intertwined; the people we interact with on a daily basis almost always live near us. As people spend more time online, data regarding these two dimensions -- geography and social relationships -- are becoming increasingly precise, allowing us to build reliable models to describe their interaction. These models have important implications in the design of location-based services, security intrusion detection, and social media supporting local communities.

Using user-supplied address data and the network of associations between members of the Facebook social network, we can directly observe and measure the relationship between geography and friendship. Using these measurements, we introduce an algorithm that predicts the location of an individual from a sparse set of located users with performance that exceeds IP-based geolocation. This algorithm is efficient and scalable, and could be run on a network containing hundreds of millions of users.

21. Ajao, O., Hong, J., & Liu, W. (2015). A survey of location inference techniques on Twitter. Journal of Information Science, 41(6), 855–864. https://doi.org/10.1177/0165551515602847

The increasing popularity of the social networking service, Twitter, has made it more involved in day-to-day communications, strengthening social relationships and information dissemination. Conversations on Twitter are now being explored as indicators within early warning systems to alert of imminent natural disasters such as earthquakes and aid prompt emergency responses to crime. Producers are privileged to have limitless access to market perception from consumer

comments on social media and microblogs. Targeted advertising can be made more effective based on user profile information such as demography, interests and location. While these applications have proven beneficial, the ability to effectively infer the location of Twitter users has even more immense value. However, accurately identifying where a message originated from or an author's location remains a challenge, thus essentially driving research in that regard. In this paper, we survey a range of techniques applied to infer the location of Twitter users from inception to state of the art. We find significant improvements over time in the granularity levels and better accuracy with results driven by refinements to algorithms and inclusion of more spatial features.

4- How can we Identify Gender on Twitter Data: Turkey Example

1. Introduction

Twitter is one of the most used social media platforms both in Turkey and the world. Twitter, whose use and importance has increased over the years, can offer us many alternatives to understand societies. In this report, we will try to identify the professions, such as how to reveal the gender criterion, which is one of the main points of evaluating users on Twitter, what should be considered while doing this, and what problems arise from the example of Turkey. The reason why it is important to conduct this research in Turkey is that there is only a little research done in both non-English and Turkish contexts. Also according to statistics, Turkey is the seventh country which has the most Twitter users in the world.

2. Methods

In gender identification from tweets, the difficulty lies in working with short text messages rather than using traditional text documents. Further, tweets are informal in their nature. Moreover, social media users tend to hide their identity, to fake gender information. Thus, gender identification from the tweets of Twitter users is a challenging problem (Sezerer, 2019, p. 204).

In order to have a balanced collection with respect to each gender, we can use common names from each gender as search filters. In the determination of common names, there are websites that suggest names to male/female babies and a name database of Turkish Language Agency (Tr. Turk Dil Kurumu). After constructing the name database, we have to eliminate names that appear on the name list of both genders and also some names that are known as unisex.

The motivation behind this is that in order to identify gender, we need active users who have enough tweets on their own, and photos are taken to supply a different type of data to help annotators in their task. After retrieving those users, they are auto labeled by their name's gender category.

3.1 Turkish language and gender inference

In one study, scholars suggest that Turkish achieved notably higher accuracy than the three other different non-English texts (Indonesian, French, Japanese), which is the highest of all four languages considered. In fact, to the best of our knowledge, this is the highest accuracy achieved in the entire Twitter gender inference literature on a dataset drawn from the Twitter general population. The k-top lists of male and female words again give some justification for the classifier's performance. Many differences between the male and female lists can be linked to men and women talking about different topics, or to different people. Several of the male words refer to soccer (gol 'goal', galatasaray 'popular Istanbul team', maç 'match', at '[part of imperative for] score'), which men plausibly tweet about more. As with Indonesian, a concern is that topics represent a biased sample of the population. Thus, we tested a classifier with soccer-specific terms removed, and again found no difference in accuracy. Many other k-top words are familiar terms of address for men (lan, abi, kardesim, adam, kanka) or a greeting used mainly between men (eyvallah), suggesting that male users are addressing or discussing men more often than female users are. In contrast, 9/25 of the k-top female words are pronouns referring to the speaker, a familiar addressee, or a third party (he/she/it), while none of the k-top male words are, suggesting female users are more often talking directly about themselves or to others. Finally, 2/25 of the k-top male words are profanity (amk, ulan), while none of the female k-top words are, suggesting male users swear more. (Sonderegger et al, 2013)

3.2 Emoji usage and gender

Concerning topics, emojis are used more frequently to communicate about issues which are perceived as trivial or less serious, and to establish or maintain social relationships. As regards the gender variable, the study confirms previous research that found a higher use of emojis by females. Gender and the expression of certain meanings through emojis also turn out to be statistically dependent variables. Expressions of love, amusement, sadness, and encouragement are gender-dependent, as are expressions of agreement and reflection.(Rua, 2021)

4. Importance of visuals and images

According to one study, in both gender and Age identification, user's profiling is much more reliable and provable with the shared photos then textual information. However, in order to define the gender variable, multimodal methods are the most appropriate way for inference.

Deep models are composed of multiple processing layers that allow to learn representations of data with multiple levels of abstraction . For instance, when an image is propagated in a pre-trained deep model, it is processed layer-by-layer transforming an array of pixel values (an image) into a representation that amplifies important aspects of the input and suppress irrelevant variations for discrimination . This methodology has reported outstanding results in a number of computer vision tasks. Our intuition is that this type of representation can be beneficial in solving the posed task.*

Images shared by social media users tend to be strongly correlated to their thematic interests as well as to their style preferences.Moreover, the image source matters (i.e. tweeted

or retweeted), and it is possible to exploit it for achieving better results on age and gender identification.

On the other hand, results indicated that images tend to be more relevant than text for determining the gender of Twitter users. Also the usefulness of visual information is somewhat dependent on the language of tweets.

5.Conclusion

We can summarize the gender identification with its positive and negative points;

Most helpful determinators

- Pronouns
- Shared photos
- Name of the people
- Enough tweets from that account
- Frequency of emoji usage (in some cases)
- Email addresses

Challenging points

- Retweets
- Uncommon using of twitter
- Other gender identities and sexual orientations (non-binary, LGBTQ+, etc.)
- Not using the real name

5- Twitter Age Identification[8]

According to the information compiled from the data in Statista, the country that uses Twitter the most in the world as of January 2021 is the United States of America with 69.3 million, while Turkey ranks 7th in the list with 13.6 million Twitter users. Young people are the most active age group (18-35 years). So what are the methods and materials used to make these analyzes?

[8] https://github.com/rumeysakars/SCQ/blob/main/CODE

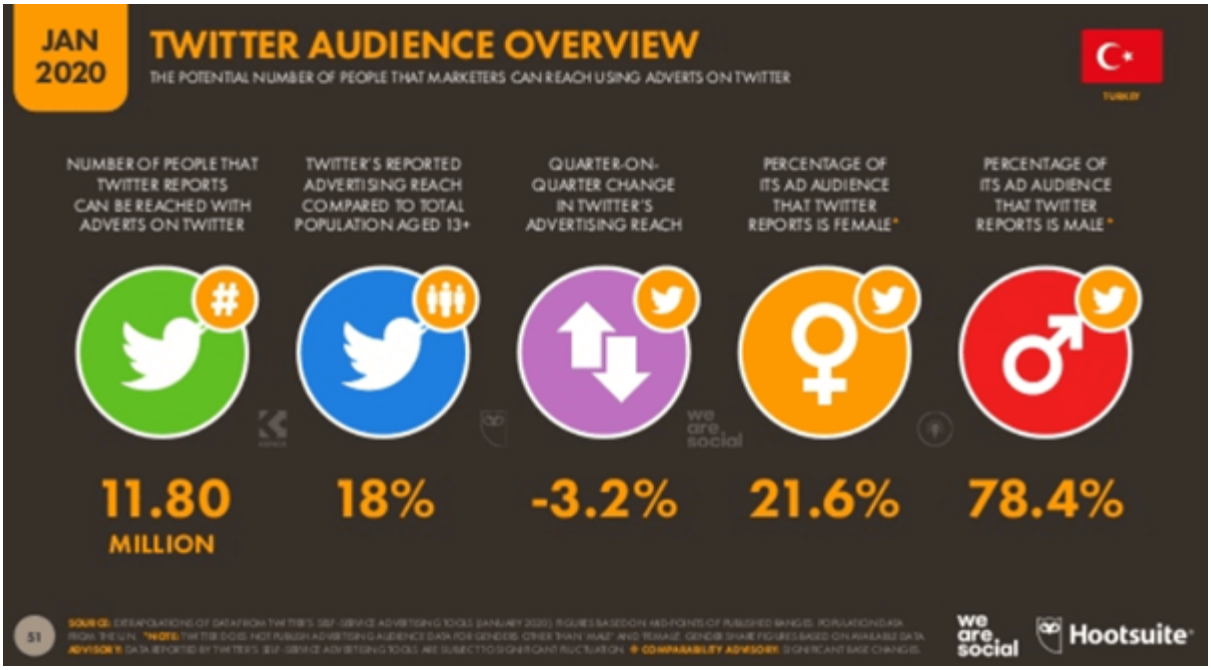

Figure 1. General summary of the Twitter Turkey audience

Although Twitter profiles do not have an age field that can be extracted via Twitter API, the age of the user can be determined from the profile description (biography) by word analysis [1]. On Twitter, words like 'age' or 'born' or 'born in' can be declared as keywords so that the ages of some of the users can be estimated. (if written in his biography)

Age was determined by considering patterns such as DD/MM/YYYY or born in YYYY. English language can be detected in this method, meaning most of the tweets (40%), this method can help to classify people in terms of age. Of course, as a disadvantage of this study, if there is no year or age information in the bio, the age information of that user will not be obtained[1] [2] .

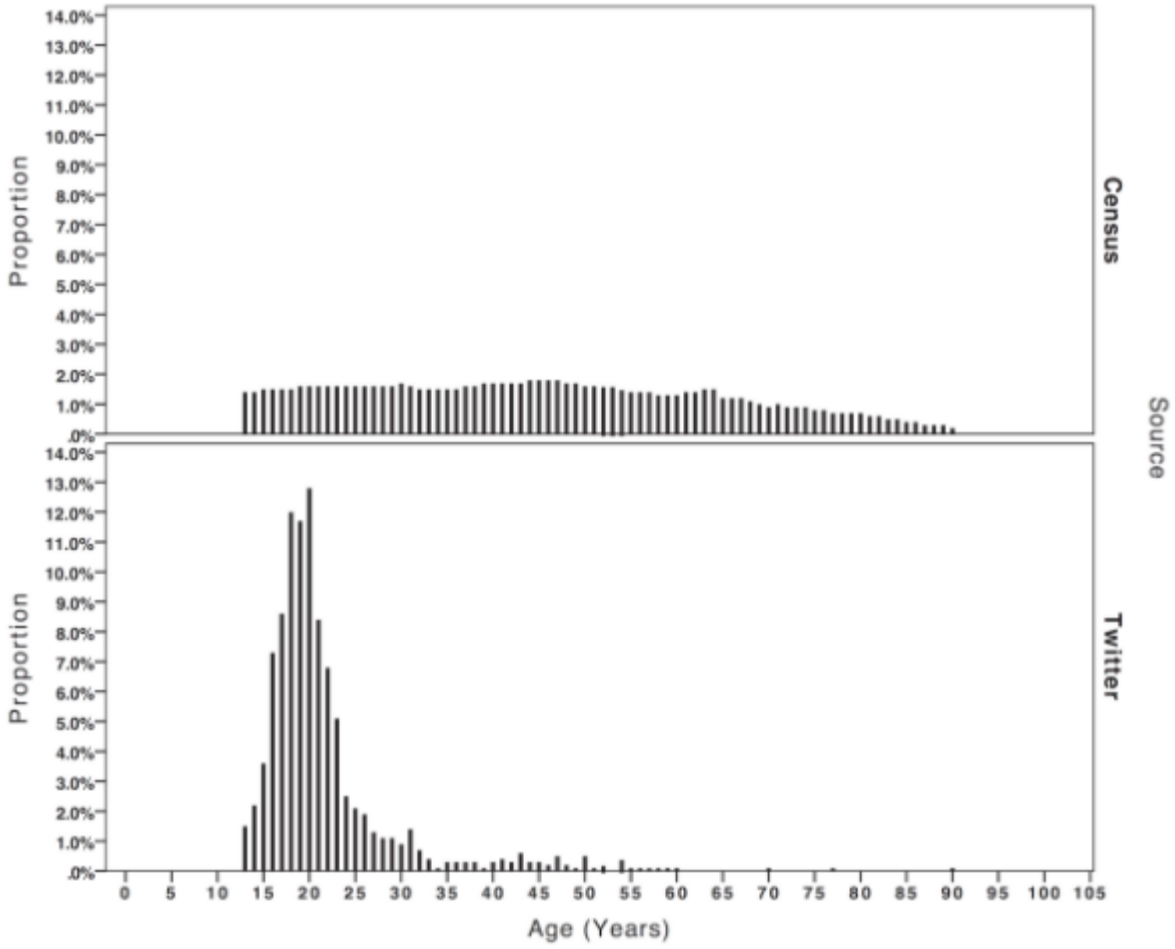

Figure 2. Detection of population and age of Twitter users

Looking back at the results of the Technology Tracker study conducted by Ipsos MORI, this study clearly shows that 93.9% of Twitter users were 35 or younger, while nearly two-thirds of Twitter users were under the age of 35 in Q3 2014.

There could be two possible reasons for this. First, the older population is less likely to state their age on Twitter. Second, the age distribution in the survey data is a function of sample bias (ie, respondents over the age of 35 who participated in the survey are particularly technology savvy). This discrepancy between emerging (traditional) and naturally occurring (new) forms of social data can be further examined.

It is obvious that young people are more crowded in terms of user rate in social media.

Age Group:	Prop. of Users:	Approx. Number of UKUsers:
13 to 20	59.4%	8,955,000
21 to 30	31.6%	4,680,000
31 to 40	4.4%	630,000
41 to 50	3.4%	510,000
51 to 60	1.1%	165,000
60+	0.3%	45,000

Figure 3. Proportion of users living in the UK and distribution by age

Another study [2] investigated the relationship between stylistic and syntactic features and the age and income of the authors. It has been found that writing style predicts income, even beyond age, for a multitude of feature types. The predictive power of writing style features was analyzed in a regression task on two datasets of approximately 5,000 Twitter users each.

Method: Both linear and non-linear machine learning regression methods were used to predict and analyze user income and age. They showed that measures of writing style showed large correlations with both age and income, and writing style predicted income even beyond age. Finally, Twitter data provides a unique opportunity to examine written variation over time

Conclusion: At the syntactic level, it was observed that increased use of nouns, adjectives and adverbs correlated more with age rather than income, while high pronoun and exclamation rate was a good predictor of low income, but only to a lesser extent with younger age.

Table 2. Obtained results [2]

Features	**Income (D_1)**		**Age (D_2)**		**Income-Age (D_1)**	
Readability	Lin	RSVM	Lin	RSVM	Lin	RSVM
ARI	.282	.311	.269	**.318**	.230	.263
Flesch-Kincaid	.285	.319	.263	.310	.234	.284
Coleman-Liau	.230	.197	.203	.265	.202	.289
Flesch RE	.277	**.345**	.186	.295	.239	**.318**
FOG	.291	.309	.222	.270	.238	.267
SMOG	.288	.339	.240	.263	.234	.301
LIX	.208	.286	.215	.268	.177	.245
ALL	**.301**	**.380**	**.278**	**.329**	**.249**	**.354**
Syntax	Lin	RSVM	Lin	RSVM	Lin	RSVM
Nouns	.155	.200	.278	**.302**	.078	**.150**
Verbs	.044	.071	(.046)	.104	.093	.114
Pronouns	.264	**.297**	.148	.180	.114	.127
Adverbs	.115	.110	.077	.111	.135	.131
Adjectives	(.030)	.149	.162	.200	(.046)	.139
Determiners	(.040)	.070	.135	.154	.103	.121
Interjections	.123	.188	.084	.122	.059	.139
ALL	**.323**	**.258**	**.319**	**.229**	**.299**	**.267**
Style	Lin	RSVM	Lin	RSVM	Lin	RSVM
Named entities	.241	**.288**	.282	.293	.255	**.281**
Contextuality	(.044)	.204	.287	**.310**	(.030)	.134
Abstract words	.108	.120	.141	.183	.125	.139
Hedging	(.019)	.079	(.015)	.000	(.000)	.083
Specific (num)	.093	.011	.072	.176	.059	.124
Elongations	.097	.160	.072	.073	.056	.114
Hapax legom.	.056	.066	.160	.219	.064	.067
ALL	**.279**	**.347**	**.306**	**.134**	**.296**	**.312**
Surface	Lin	RSVM	Lin	RSVM	Lin	RSVM
# char. / token	.085	.144	.104	.148	.051	.101
# tokens / tweet	.158	.159	.228	.237	.115	.116
# char. / tweet	.214	**.261**	.262	**.278**	.153	**.169**
# words >5 char.	.139	.191	(.009)	.087	.112	.163
Type/token ratio	.099	.132	.090	.180	.100	.126
Punctuation	.218	.123	.093	.086	.057	.084
Smileys	.064	.113	.146	.144	(.030)	.090
URLs	.084	.128	.187	.194	(.040)	.077
ALL	**.379**	**.330**	**.294**	**.307**	**.352**	**.126**

Another study [3] investigated age estimates based on tweets from Twitter users, focusing on the link between age and language use. Here's how we can explore age estimation: classify users by age categories based on life stages and estimate their exact age. It was found that an automated system outperformed humans in these tasks, and both humans and automated systems had difficulty estimating the age of the elderly.

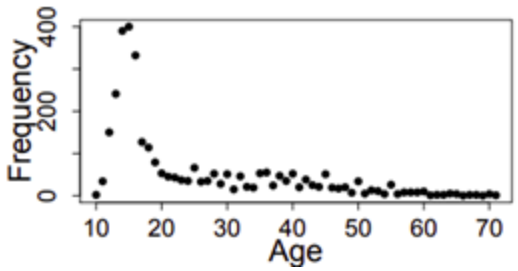

Figure 4. Number of tweets by age

As discussed earlier in this article, it may be more natural to divide users by life stages rather than a fixed age category. Life stages can be approached from different dimensions. In this article, students, employees, retirees, etc. By making a distinction, the life stages of the people were used according to their occupations.

Table 3. Life stage and tweet frequency

Life Stage	Frequency
Secondary school student	1352
College student	316
Employee	1021
Retired	5
Other	15
Unknown	132
Not accessible	344

Pearson correlation coefficients between variables and actual age were calculated using the same data from age estimation experiments. Younger people have been found to use more obvious stylistic changes, such as alphabetic extension and capitalization of words. Older people tend to use more complex language with longer tweets, longer words, and more prepositions. Older people also have more links and hashtag usage, which can be associated with information sharing and impression management. Surprisingly, the use of pronouns is one of the most studied variables regarding age.

The system was able to predict the exact age within a margin of 4 years. It was also found that most changes occur when people are young, and that the variables studied after about 30 years show little change. This may also explain why it is more difficult to estimate the age of the elderly.

Works with data (JSON file)

In the json file (Total Merge userobjects_01.07.2021 - Ozgem.json), age were investigated by finding some words. With the word “born”,“born in”, “years old”,”yas”,”yasi”, “19” there was no match with the file.

“Happy n^{th} birthday” can be used to identify the age of user. However, in the data file, there is no “mutlu” or “yillar” words if it is translated to Turkish language. [4]

The source and article [5][6] will be examined as well.

1287 users have been detected as under eighteen.

324 users have been detected between 19 age and 23 age.

3160 have been detected between 23 age and 40 age.

296 have been detected between 40 age and 60 age.

5067 number of users’ age have been identified.

Features that can identify the age

The first is that the older population is less likely to state their age on Twitter. The second is that the age distribution in the survey data is a function of sample bias (i.e. participants over the age of 35 in the survey were particularly tech-savvy).

The fact that younger people are more populous on social media as a proportion of the user population has been indicated by previous studies, but even the 1.1% of users between the ages of 51 and 60 could account for approximately 165,000 people in the UK. Taking this further and making the assumption that the age profile of users across the world is similar to the UK, they estimate the worldwide number of users between 51 and 60 years to be in the region of 2,981,000. Even this is a conservative estimate as it is calculated based 271,000,000 users who are active every month. (Sloan L, Morgan J, Burnap P, Williams M (2015) Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data. PLoS ONE 10(3): e0115545. doi:10.1371/journal.pone.0115545)

Text style can identify the age of user

Other linguistic variables were considered based on popular Internet conventions, such as use of excessive capitalization or punctuation (e.g., “WHAT!?!?!”), alphabetical lengthening (e.g., “that was sickkkkk”), use of emojis, and acronyms with Internet origins (e.g., “lol,” “omg”) Also, the age of twitter account

To model age, they tested six different classifiers (logistic regression, support vector machines, random forests, adaBoost, and extra trees) and included a dummy classifier to assess baseline performance.

1. I am X years old 2. Born in X 3. X years old phrases has been used. In our data, this approach was not proper. (Sloan L, Morgan J, Burnap P, Williams M (2015) Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data. PLoS ONE 10(3): e0115545. doi:10.1371/journal.pone.0115545)

Network can identify the age of user

They might also hypothesis that young people are more likely to profess their age in their profile data and that this would lead to an overestimation of the ‘youthfulness’ of the UK Twitter population. As this is a new and developing field we have no evidence to support this claim, but the following discussion and estimations should be treated cautiously. What follows is their best attempt to profile the age distribution on Twitter in the absence of other baseline information.

(Sloan L, Morgan J, Burnap P, Williams M (2015) Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data. PLoS ONE 10(3): e0115545. doi:10.1371/journal.pone.0115545)

In this paper, we devise a language-independent methodology for determining the age of Twitter users from data that is native to the Twitter ecosystem. The key idea is to use a Bayesian framework to generalise ground-truth age information from a few Twitter users to the entire network based on what/whom they follow. Their model performs as well as the current state of the art for inferring the age of Twitter users without being limited to specific linguistic or engineered features. They have successfully applied the model to infer the age of 700 million Twitter users demonstrating the scalability of our approach.(Probabilistic Inference of Twitter Users' Age based on What They Follow-arXiv:1601.04621)

https://blog.twitter.com/en_us/a/2016/age-data-is-now-available-from-the-gnip-audience-api

<12	12–13	14–15	16–17	18–24	25–34	35–44	45–64[5]	65+
vlogger	child presenter	child singer	singer	metalcore band	hip hop duo	hip hop artist	evangelist	political journalist
minecraft gamer	YouTuber	child singer	metalcore band	rock band	boy band	rapper	evangelist	retired cyclist
internet personality	child actress	child singer	deathcore singer	rapper	boy band	history channel	evangelist	golf channel
vlogger	child actress	child singer	electronic band	computer game	comedian	record label	faith group	retired rugby player
gaming commentator	girl band	child singer	electronic band	rock band	adult actress	boxer	faith magazine	boxer

Figure 1. Ages and their following interests

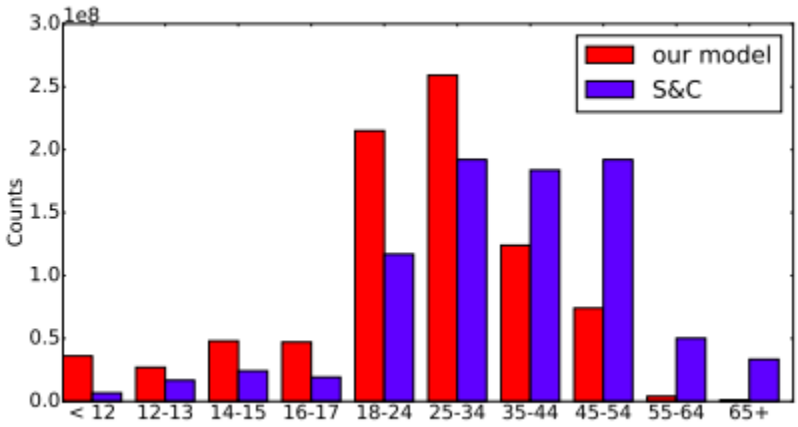

Figure 2. Their model and the survey data and comparison of them

Job can identify the age of user

No information about it.

Photo can identify the age of user

https://help.twitter.com/en/safety-and-security/age-verification

I could not understand what they are trying to do.

Name can identify the age of user

No info about it.

The date of having the account can identify the age of user

https://books.google.com.tr/books?id=PoBSDwAAQBAJ&pg=PA386&lpg=PA386&dq=text+style+twitter+age+identification&source=bl&ots=FFBXPvBVwC&sig=ACfU3U2NHtCSSiwq7DrQu-Ao

eFj2wovp4g&hl=tr&sa=X&ved=2ahUKEwirprLZg8_zAhU0gP0HHSg_BrcQ6AF6BAgCEAM#v=onepage&q=text%20style%20twitter%20age%20identification&f=false

https://www.researchgate.net/post/How-can-I-get-users-age-from-Twitter

https://stackoverflow.com/questions/29043138/using-tweepy-to-determine-the-age-on-an-account

6- Interpretive Quantitative Sociology

According to Babones (2016), sociological variables are observed manifestations of **unobserved** actions and processes. It is a different approach from positivism that takes for granted variables causes variables. **Interpretive quantitative research** uses statistics to illuminate the unobservable data generating processes that underlie observed data.

This kind of research prioritizes **discovery** over testing through the data-generating process. Implying this approach to the available data is possible through **multiple variable operationalizations** and **multiple statistical models**. The probability of causality between concepts and ways where those concepts might be measured are limitless.[1] As a result, estimating what lies unobserved beneath is possible by using a **triangulation** tool that helps to compare data from multiple sources at multiple levels. For example, triangulation could be applied while looking at differences in the relationship between **education** and **income** across different **subgroups** in society.

However, triangulation is not sufficient by itself; researchers also need to embrace **reflexivity** to improve understanding of society. Reflexivity is a process that researchers examine their own personal roles in research they conduct (Alvesson & Sköldberg, 2009). Reflexive research could be implied to quantitative methods through Ryan and Golden's (2006) dialectical formulation among:

a) the researcher's constructs of the concepts that they measure using observed variables,
b) the commonsense meanings of those variables,
c) the research data,
d) the researcher's ideological biases and
e) the structural and historical forces that underlie the research.

In conclusion, because society is constantly changing, positivist approaches that depend on a priori theorization may not help understand society. Interpretive research leads to uncover society's dynamics where the observed correlation between variables rises or falls (Babones, 2016).

7- Detecting Ethnicity/Race from Twitter

Author(s)	Year	Article	Method
Mislove, Alan & Lehmann, Sune & Ahn, Yong-Yeol & Onnela, Jukka-Pekka & Rosenquist, (James)	2011	Understanding the Demographics of Twitter Users	— Detecting race/ethnicity using self-reported last names — Correlate the last name with data from the U.S. 2000 Census — E.g., "Myers" --> 86% Caucasian, 9.7% African-American…
Huang, Wenyi & Weber, Ingmar & Vieweg, Sarah	2014	Inferring nationalities of Twitter users and studying inter-national linking	— Used crowdsourced data (CrowdFlower) for data labeling and validation — Features: Location-based features (followers' locations, friends' locations, self-stated locations, geo-tagged tweets' locations), time zone, language related features, hashtags, profile picture features (Faceplusplus; gender, race & age), name ethnicity (name ethnicity detection toolkit), UTF-8 charset type, tweet source, mentioned users — Gradient Boosted Tree
Olteanu, A., Weber, I., & Gatica-Perez, D.	2015	Characterizing the Demographics Behind the #BlackLivesMatter Movement	— No prediction prediction/detection of race/ethnicity/demographics — Crowdsourced annotation (CrowdFlower)
Bokányi, E., Kondor, D., Dobos, L., Sebők, T., Stéger, J., Csabai, I., & Vattay, G.	2016	Race, religion and the city: twitter word frequency patterns reveal dominant demographic dimensions in the United States	— Unsupervised; geotagged tweets; combination of Latent Semantic Analysis (LSA) & Robust Principal Component Analysis (RPCA) — County-level aggregation --> comparison with census data — Bag-of-words; word-frequencies. Normalized by the total number of words posted for each county; inverse document frequency weighing — Conclusions: — Geographical closeness implies closeness in the semantic space — The most important factor behind the variation in the language use of different

			counties is the presence of Afro-American ethnicity, followed by population density
Cesare, N., Grant, C., & Nsoesie, E. O.	2017	Detection of User Demographics on Social Media: A Review of Methods and Recommendations for Best Practices	— Methods to detect race and ethnicity: Supervised learning approaches (8 papers), adjusted data matching (2 papers), facial recognition (2 papers) user names, posts, or both (4 papers) — Features that are a good indicator of race or ethnicity: Profile photos, user profile descriptions, user location — Challenges: unavailability of ground truth data, inconsistency in the performance metrics among different studies
Messias, J., Vikatos, P., & Benevenuto, F.	2017	White, Man, and Highly Followed: Gender and Race Inequalities in Twitter	— Used Face++, a face recognition software able to recognize gender and race of identifiable faces in the user's profile pictures — Methods mentioned in the "Related Work" section: Self-descriptions, profile images (Face++), users' names, tweets, social networks, hashtags
Daniel Preoţiuc-Pietro and Lyle Ungar	2018	User-Level Race and Ethnicity Predictors from Twitter Text	— Used a dataset of users who self-report their race/ethnicity through a survey (through Qualtrics) — Developed predictive models from text — Features: Unigrams (bag-of-words), LIWC (Linguistic Inquiry Word Count), Word2Vec Topics, Sentiment & Emotions, Part-of-Speech Tags, Linear Ensemble
Wood-Doughty, Z., Xu, P., Liu, X., & Dredze, M.	2021	Using Noisy Self-Reports to Predict Twitter User Demographics	— Used a method to identify self-reports of race and ethnicity from Twitter profile descriptions — Keyword-matching: — Racial keyword: black, african-american, white, caucasian, asian, hispanic, latin, latina, latino, latinx — Person keyword: man, woman, person, individual, guy, gal, boy, girl

Conclusion

References

Sang, E. T. K., & Van den Bosch, A. (2013). Dealing with big data: The case of twitter. *Computational Linguistics in the Netherlands Journal*, *3*, 121-134.

1- Profiling Political Behavior in Turkey

Newton, K., & Giebler, H. (2008). Patterns of participation: Political and social participation in 22 nations. *Patterns of Participation: Political and Social Participation in 22 Nations*.

https://www.econstor.eu/bitstream/10419/49726/1/587500093.pdf.

https://libguides.princeton.edu/politics/indicators

http://countrystudies.us/turkey/85.htm

http://www.mustafaaydin.gen.tr/20/turkiye-sosyal-siyasal-egilimler-political-and-social-trends-in-turkey

https://www.mei.edu/multimedia/podcast/turkeys-social-and-political-trends

https://www.khas.edu.tr/tr/node/4909

https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199604456.001.0001/oxfordhb-9780199604456-e-017

https://www.jstor.org/stable/j.ctt7zvbrd

https://eacea.ec.europa.eu/national-policies/eurydice/content/political-social-and-economic-background-and-trends-103_en

https://santandertrade.com/en/portal/analyse-markets/turkey/economic-political-outline?url_de_la_page=%2Fen%2Fportal%2Fanalyse-markets%2Fturkey%2Feconomic-political-outline&&actualiser_id_banque=oui&id_banque=0&memoriser_choix=memoriser.

http://countrystudies.us/turkey/85.htm

Voting Behavior of Turkish People References:

https://konda.com.tr/wp-content/uploads/2018/05/KONDA_SecmenKumeleri_AkParti_Secmenleri_Mayis2018.pdf

https://www.cumhuriyet.com.tr/galeri/son-anket-cumhur-ittifakinin-oylari-eridi-1849695

https://sahamerkezi.org/turkiye-geneli-siyasal-gundem-ve-secmen-egilim-anket-calisma-raporu/

https://sonararastirma.com.tr/turkiye-geneli-siyasi-egilimler-ve-gundem-arastirmalari/

https://www.asarda.com/anketler/turkiye-secmen-egilim-arastirmasi-2020/

https://biruni.tuik.gov.tr/secimdagitimapp/halksecmen.zul?

Country Profiles References:

https://guides.library.yale.edu/c.php?g=595576&p=4297809

https://www.oecd-ilibrary.org/economics/country-statistical-profiles-key-tables-from-oecd_20752288

https://www.oecd.org/regional/regional-policy/country-profiles.htm

https://www.bbc.com/news/world-europe-17988453

https://www.hrw.org/world-report/2020/country-chapters/turkey#

http://unctadstat.unctad.org/countryprofile/GeneralProfile/en-GB/792/index.html

https://eacea.ec.europa.eu/national-policies/eurydice/content/political-social-and-economic-background-and-trends-10_3_en

2- Twitter User Types

Chu, Zi & Gianvecchio, Steven & Wang, Haining & Jajodia, Sushil. (2012). Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg?. Dependable and Secure Computing, IEEE Transactions on. 9. 811-824. 10.1109/TDSC.2012.75.

de Silva, L., & Riloff, E. (2014). User Type Classification of Tweets with Implications for Event Recognition. *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*. Published. https://doi.org/10.3115/v1/w14-2714

S. Yardi, D. Romero, G. Schoenebeck, and D. Boyd, “Detecting Spam in a Twitter Network,” First Monday, vol. 15, no. 1, Jan. 2010.

Uddin, M.M., Imran, M., & Sajjad, H. (2014). Understanding Types of Users on Twitter. ArXiv, abs/1406.1335.

3- Location Identification

Foucaud, Florent & Mertzios, George & Naserasr, Reza & Parreau, Aline & Valicov, Petru. (2014). Identification, location-domination and metric dimension on interval and permutation graphs. I. Bounds. Theoretical Computer Science. 668. 10.1016/j.tcs.2017.01.006.

Chappell, P., Tse, M., Zhang, M., & Moore, S. (2017). Using GPS geo-tagged social media data and geodemographics to investigate social differences: A twitter pilot study. *Sociological Research Online, 22*(3), 38-56. https://doi.org/10.1177/1360780417724065

Burrows, R., Webber, R., & Atkinson, R. (2017). Welcome to ‘Pikettyville’? Mapping London’s alpha territories. The Sociological Review, 65(2), 184–201. https://doi.org/10.1111/1467-954X.12375

Graham, M., Hale, S. A., & Gaffney, D. (2014). Where in the world are you? Geolocation and language identification in Twitter. *The Professional Geographer*, *66*(4), 568-578.

Graham, S. D. (2005). Software-sorted geographies. *Progress in human geography*, *29*(5), 562-580.

Harris, R., Sleight, P., & Webber, R. (2005). *Geodemographics, GIS and neighbourhood targeting* (Vol. 8). John Wiley & Sons.

Parker, S., Uprichard, E., & Burrows, R. (2007). Class places and place classes geodemographics and the spatialization of class. *Information, Communication & Society*, *10*(6), 902-921.

Sloan, L., & Morgan, J. (2015). Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. *PloS one*, *10*(11), e0142209.

Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P., & Rana, O. (2013). Knowing the Tweeters: Deriving sociologically relevant demographics from Twitter. *Sociological research online*, *18*(3), 74-84.

Uprichard, E., Burrows, R., & Parker, S. (2009). Geodemographic code and the production of space. *Environment and Planning A*, *41*(12), 2823-2835.

Webber, R., & Butler, T. (2007). Classifying pupils by where they live: how well does this predict variations in their GCSE results?. *Urban Studies*, *44*(7), 1229-1253.

Mourad A., Scholer F., Sanderson M. (2017) Language Influences on Tweeter Geolocation. In: Jose J. et al. (eds) Advances in Information Retrieval. ECIR 2017. Lecture Notes in Computer Science, vol 10193. Springer, Cham. https://doi.org/10.1007/978-3-319-56608-5_26

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010, April). Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web* (pp. 851-860).

Sadilek, A., Kautz, H., & Bigham, J. P. (2012, February). Finding your friends and following them to where you are. In *Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 723-732).

Roller, S., Speriosu, M., Rallapalli, S., Wing, B., & Baldridge, J. (2012, July). Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 1500-1510).

Priedhorsky, R., Culotta, A., & Del Valle, S. Y. (2014, February). Inferring the origin locations of tweets with quantitative confidence. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 1523-1536).

Kinsella, S., Murdock, V., & O'Hare, N. (2011, October). " I'm eating a sandwich in Glasgow" modeling locations with tweets. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents* (pp. 61-68).

Jurgens, D., Finethy, T., McCorriston, J., Xu, Y. T., & Ruths, D. (2015, April). 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*.

Cheng, Z., Caverlee, J., & Lee, K. (2010, October). You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 759-768).

Backstrom, L., Sun, E., & Marlow, C. (2010, April). Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web* (pp. 61-70).

Ajao, O., Hong, J., & Liu, W. (2015). A survey of location inference techniques on Twitter. Journal of Information Science, 41(6), 855–864. https://doi.org/10.1177/0165551515602847

4- How can we Identify Gender on Twitter Data: Turkey Example

Erhan Sezerer, Ozan Polatbilek, Selma Tekir,(2019), "A Turkish Dataset for Gender Identification of Twitter Users" Proceedings of the 13th Linguistic Annotation Workshop, pages 203–207 Florence, Italy, August 1, 2019.Association for Computational Linguistics

"Gender Inference of Twitter Users in Non-English Contexts", (2013) Sonderegger Morgan, Ruths, Derek Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1136–1145, Seattle, Washington, USA, 18-21 October 2013 Association for Computational Linguistics

López-Rúa, P. (2021). Men and women on Twitter: A preliminary account of British emoji usage in terms of preferred topics and gender-related habits. Language@Internet, 19, article 3. (urn:nbn:de:0009-7-52418)

*Alvarez Carmona, Miguel A., Pellegrin Luis,(2018). " A visual approach for age and gender identification on Twitter", Instituto Nacional de Astrofísica, Óptica y Electrónic Luis Enrique Erro 1, Puebla 72840, México

https://gender-api.com

5- Twitter Age Identification

[1] Article Source: Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data

Sloan L, Morgan J, Burnap P, Williams M (2015) Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data. PLOS ONE 10(3): e0115545. https://doi.org/10.1371/journal.pone.0115545

[2] Flekova, L., Preotiuc-Pietro, D., & Ungar, L. (2016). Exploring Stylistic Variation with Age and Income on Twitter. ACL.

[3] Nguyen, D., Gravel, R., Trieschnigg, D., & Meder, T. (2013). "How Old Do You Think I Am?" A Study of Language and Age in Twitter. In Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013. (pp. 857–862). The AAAI Press. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/5984

[4] Morgan-Lopez AA, Kim AE, Chew RF, Ruddle P (2017) Predicting age groups of Twitter users based on language and metadata features. PLoS ONE 12(8): e0183537.

[5] http://m3.euagendas.org/

[6] Wang, Z., Hale, S. A., Adelani, D., Grabowicz, P. A., Hartmann, T., Flöck, F., & Jurgens, D. (2019)

6- Interpretive Quantitative Sociology

Alvesson, M. and Sköldberg, K. (2009). Reflexive Methodology: New Vistas for Qualitative Research. London: SAGE.

Babones, S. (2016). Interpretive Quantitative Methods for the Social Sciences. Sociology, 50(3), 453–469. https://doi.org/10.1177/0038038515583637

Ryan, L. and Golden, A. (2006). 'Tick the box please': A reflexive approach to doing quantitative social research. Sociology 40: 1191–200.