# City-level User Location Detection from Social Media and Analysis of Political Participation
# Sosyal Medya Kullanıcı Verilerinden Seçmen Şehir Bilgisi Tespit Etme ve Siyasi Katılım Analizi

2024 32nd Signal Processing and Communications Applications Conference (SIU) | 979-8-3503-8896-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/SIU61531.2024.10600825

Alp Önder Yener
Sabanci University
alp.yener@sabanciuniv.edu

Ali Najafi
Sabanci University
ali.najafi@sabanciuniv.edu

Onur Varol
Sabanci University
onur.varol@sabanciuniv.edu

***Abstract*—The social media platform X (Twitter), regularly used by hundreds of thousands of users in Türkiye, serves as a space where users share content, express opinions on current issues, and interact with each other. Some of the agenda on social media reflects events that occurred on the same day, allowing users to follow closely events happening in their vicinity through the platform. The locations of these users in different cities of Türkiye can be inferred from the metadata of the tweets, but there is currently no system specifically designed for Türkiye. We first present an approach to automatically infer the user locations at the city-level. Subsequently, the validation of these inferences examined by analyzing volume of activities obtained when a politician visiting a particular city. Finally, the relationship between voting percentages at the election and changes observed in the quantity of social media posts is investigated.**

***Keywords*—Twitter, data science, reverse geocoding, elections.**

***Özetçe* —Sosyal medya platformu olan X (Twitter) Türkiyede düzenli olarak yüz binlerce kullanıcı tarafından kullanılmaktadır ve bu kullanıcılar paylaştıkları içerikler ile gündeme yönelik söylemlerde bulunmakta ve birbirleri ile etkileşmektedir. Sosyal medyadaki gündemin bir kısmı o gün gerçekleşen olayların bir yansıması olabilmekte ve kullanıcılar kendilerine yakın gerekleşen olayları da platform üzerinden takip edebilmektedir. Bu kullanıcıların Türkiye'nin hangi illerinde konumlandıkları ise paylaşılan tweetlerin metaverisinden çıkarılabilir fakat bunu bulmaya yarayan bir sistem Türkiye için bulunmamaktadır. Bu çalışmada öncelikle kullanıcının attığı tweetin metaverisinden kullanıcının bulunduğu ili otomatik olarak bulmaya yarayan bir yaklaşım sunulmakta, sonrası ise bu çıkarımların siyasilerin yaptığı seçim meetinglerine bakılarak doğrulanması amaçlanmıştır. Seçim sonucu oy oranları ve sosyal medyada gözlenen paylaşımların miktarındaki değişimler arasındaki ilişki incelenmiştir.**

***Anahtar Kelimeler*—Twitter, veri bilimi, coğrafi referanslama**

## I. Introduction

Social media have been a critical platform for political communication and campaigns [1], [2]. Wide-spread use of these platforms across the world position them as a sensor to real-world activities and people share information almost real-time on these platforms on various issues like societal events and natural disasters [3]–[7].

**979-8-3503-8896-1/24/$31.00 ©2024 IEEE**

Large-scale activity and the metadata shared along with the content can help making reliable estimates on political orientation, gender, belief, and many other personal details [8], [9]. People can also share these information willingly and others can be collected through applications such as GPS coordinates, device type, and IP address. Some of these information can be shared to 3rd-parties to improve user experience.

On Twitter – recently rebranded as X – accounts can share their profile pictures, websites and locations in their profile and coordinates can be also attached to tweets. Some of these details are available as free-form text such as location details in the profile. In this work, we use the location information shared by users to estimate their cities in Türkiye. To achieve this we process profile details and posts and to validate this estimation approach by comparing with population statistics and real-world societal events. We use political campaign events held in different cities during the 2023 Turkish Presidential Elections.

## II. Related Work

Efforts to extract useful information from social media data can be observed in large-range of applications from finance and mental health to politics [10], [11]. Application for location estimation also have significant use in advertisement and recommendation systems [12], [13]. One particular domain of interest is politics to collect insights about citizens needs and their preferences for effective campaigns [14]–[16].

Previous research efforts utilize social network information to estimate location of accounts [17] or using machine learning to extract trends from textual information [18]. Researchers also use Turkish social media posts to predict user location for three major cities (Ankara, Izmir and Istanbul) with 54.47% accuracy [19]. Social media posts can also be used for monitoring how and where the important societal events unfold [20]–[23]. Especially the literature on social protests tracks activities of users longitudinally.

## III. Dataset

In this study, we utilize a publicly available dataset on the 2023 Turkish presidential election [24], which captures political activities between March 17th and May 15th, 2023. In this dataset, we identified almost 990 thousand unique users and over 17 million tweets posted by filtering out tweets based on the mentions of the four election candidates.

For the validation analysis, we obtained population statistics and names of the counties and districts for different cities through online governmental websites [25], [26]. Since our research investigates political campaign meeting in different cities, we also collect date and location details of these events from political party websites, politicians' social media accounts, and conducting online search on news and YouTube manually. Figure 1 demonstrates the routes for the campaign meetings of four election candidates. Kemal Kılıçdaroğlu and Sinan Oğan had the most and the least number of meetings compared to other candidates, respectively. President Erdoğan's campaign activities primarily carried out by the ministers and other party representatives. Since we only focused on campaign activities of the leaders, meetings associated with Erdoğan is less than others candidates as a result.

## IV. Method

### A. Preprocessing

In our dataset, we filtered retweets to remove duplicate text and we also want to focus on original content which can suggest the location of the tweeting user. Social media text known to be noisy and can contain ambiguities due to language-specific characters. Since some users choose to tweet using these characters and others can replace them with letters from standard keyboards, we normalized the text by lowercasing and converting all characters by unidecoding.

### B. City-level location detection

To estimate the location of a user, the tokens and names of various districts in Türkiye are used to match in the user metadata. The search is carried out from the largest units (cities) to smaller one if there is not match. Later, we used the mappings of counties and districts to find the corresponding city of the county (in Turkish "ilçe") or district (in Turkish "semt") that we prepared. In Türkiye, some cities share the same district and county names. For example, there are two counties with the same name, "Edremit," in two different cities (Balıkesir and Van). When we encountered an ambiguity we picked the city with the largest population. Although this approach can make the results biased toward the popular cities, it is a reasonable heuristic considering the high number of tweets from those locations and improves results for smaller cities.

The aforementioned preprocessing pipeline was for textual metadata information. We can also capture coordinates shared by accounts which is available for data collected with Twitter API V1. These coordinated mapped to their corresponding Turkish cities using the boundaries that is available on Github[1].

## V. Findings

### A. Estimating population distribution

To validate the location estimation of our approach, we relied on the population statistics collected for Türkiye. We can compare the fraction of individuals registered these cities with the accounts we identified from each location. In Fig.2, we show association between the fractions of residents and estimated number of social media users. We considered two different version of the estimation approach: i) exact mapping with the cities and ii) predictions biased toward crowded cities when there is an ambiguity. In our comparison between these two approaches, we do not observe any significant difference. Both approaches results with a significant Spearman's correlation as shown in Fig.2.

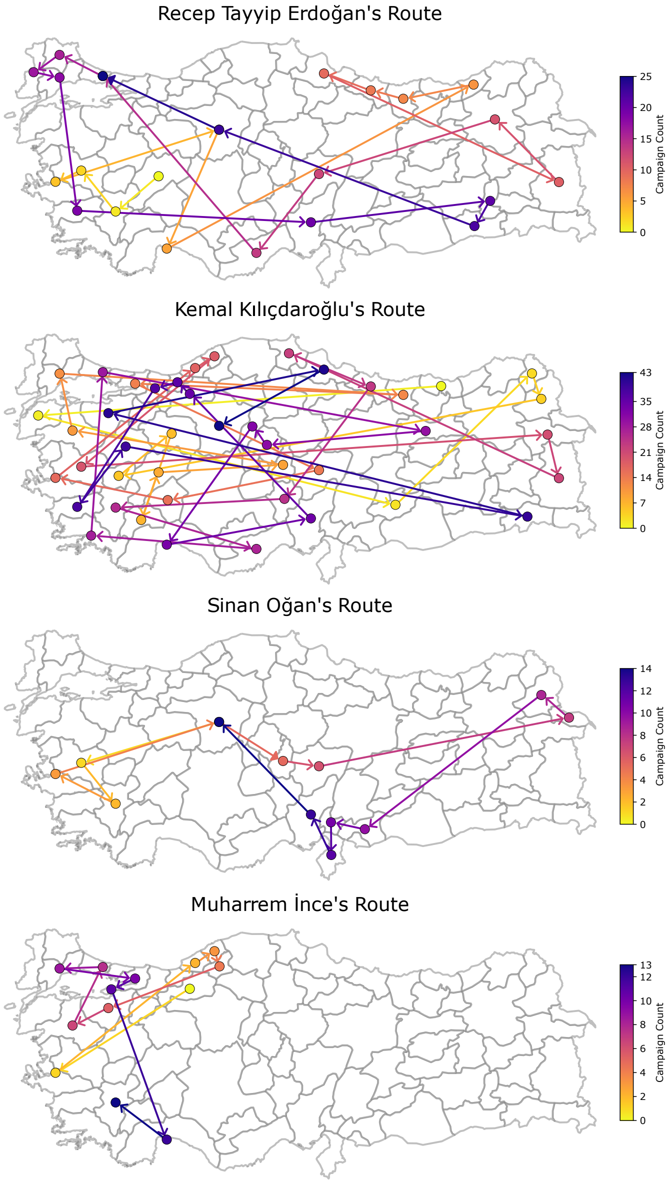

Figure 1. **Political campaign routes of presidential candidates.** Political candidates and their campaigns routes are mapped in chronological orders as also indicated by the color. Directed arrows points consecutive cities along the campaign routes. The cities that were visited twice by a candidate are colored according to the second visit.

### B. Validation with political campaign events

Detecting users location using social media data is an important challenge to monitor online activities; however, this task has very limited ground-truth datasets in general. In the case of Türkiye, we have to address this challenge to study political issues by location and our goal is to develop a

[1] https://github.com/alpers/Turkey-Maps-GeoJSON

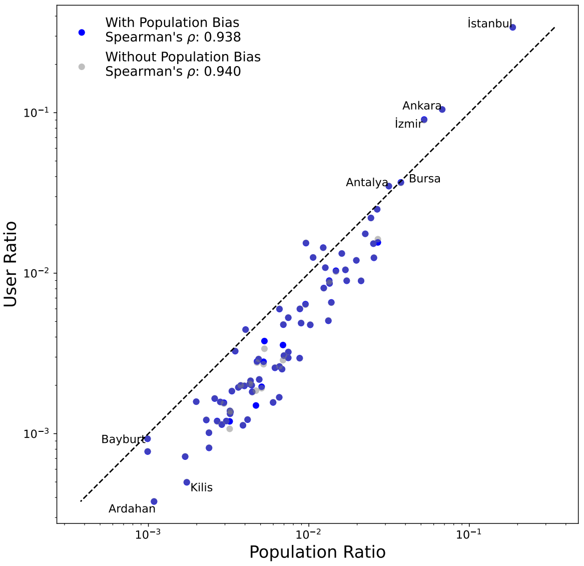

Figure 2. **Population detection.** Unique accounts of social media users on a large-scale dataset can be used to estimate population size for cities. We compared the ratio of accounts identified in each city with their population.

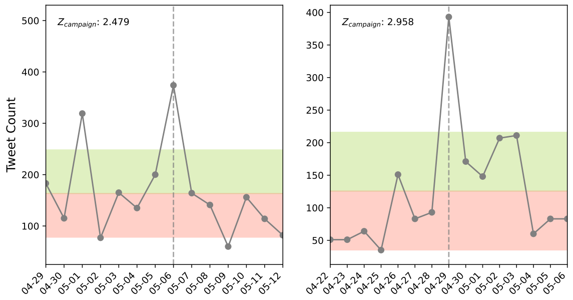

Figure 3. **Time series of online activities centered around a campaign.** Daily activities mentioning a candidate on a particular city presented as time series, which is centered around the campaign day. Z-score for the activity is calculated with respect to these two-week activity periods. Shaded areas indicate level of activities with a one standard deviation from the mean.

technique that prioritizes accuracy over recall, so the users that we identified in each location will be a sample. Since there is no ground-truth exists for this task, we utilize political campaign events as a proxy to real-world activities.

To validate the location estimation, we made a reasonable assumption: "When a candidate rally in a city, the citizens of that city will tweet more than usual about that candidate." We consider the online activities generated for each city by its users around the campaign dates and measure activity time series as shown for some example cities in Fig.3. Using the activity time series, we calculated the z-score for the activity on campaign day considering the two weeks period around the campaign meeting. Amount of activities mostly exceeds the expected values when compared with historical records.

In Fig.4, we present the z-scores for online activities in different cities. Distribution of z-scores shows bias towards

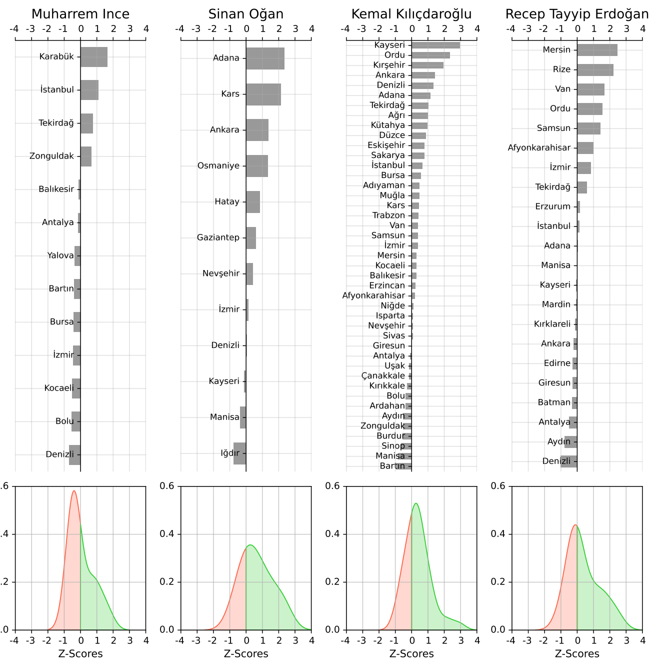

Figure 4. **Impact of political campaign meetings.** Social media users response to campaign events by posting messages online. We measure timeseries of activities on a particular city for two weeks period around a meeting. Activity on meeting days compared to other days and we measure z-score for that day. Politicians tend to receive significantly more tweet on most cities when they campaign there.

positive values as expected. The higher z-scores indicate more online reaction to that candidate's meeting. For instance, Kayseri and Mersin are the most reactive cities for Kemal Kılıçdaroğlu and Recep Tayyip Erdoğan, respectively. This approach suggest that our assumption is valid and our methodology can also identify users' location for applications on this dataset. We also want to point that meetings can have effect lasting longer than a day, crowded cities can talk about politicians regardless of their visit, and other spill-over effects from neighboring cities or multiple events taking place on the same day. Considering these additional factors, z-scores we calculated may be an under estimation of the real impact of these political meetings.

### C. Comparison of election outcomes with online campaign interactions

In 2023 elections, three candidates competed (M. İnce withdrawn 3 days prior to election), since neither of the candidates received more than 50% of the votes the election finalized in the second round. We analyzed the vote differences of the two candidates and compared it with the z-scores obtained in these locations to investigate association between the election outcomes and online reaction towards these candidates.

In Fig.5, we present an association between the vote differences and z-scores measured in different campaign locations. We measured a negative association between these quantities, meaning that candidates received more vote advantage in the location where their opponents received more surprise measured by online activities. For instance, Tayyip Erdoğan significantly won in Kayseri and citizens of that city reacts more as measured by number of tweets to Kemal Kılıçdaroğlu.

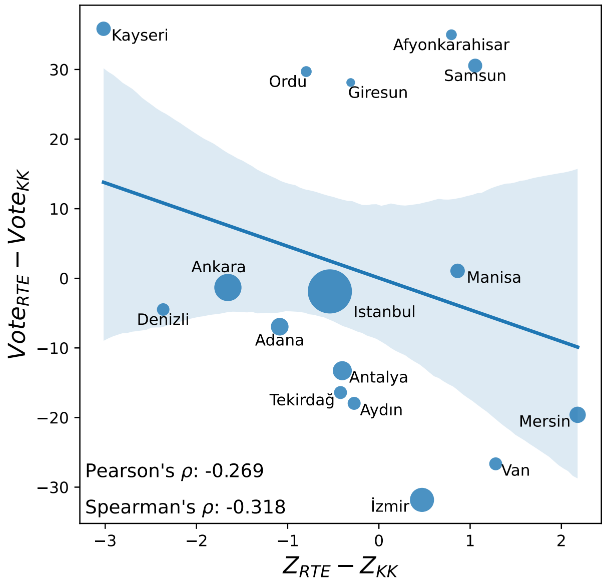

Figure 5. **Vote difference compared to user reaction.** We measure z-scores from political activities of Kemal Kılıçdaroğlu (KK) and Recep Tayyip Erdoğan (RTE) and compared the activity difference with election outcomes in the first round. Association between these entities better presented with a regression line and the 90% confidence intervals.

Similarly in Mersin, Kemal Kılıçdaroğlu was the winner, while people from Mersin send excdeedingly higher amounts to Tayyip Erdoğan. These findings suggest, voter tends to post content about their favourite candidates in general resulting with z-score around zero on campaign dates and their opponents receive more reaction than usual as a result of the external effects and anticipation of their visits.

## VI. Conclusion and Discussion

We offer an approach to estimate user location to track political participation. We validated our findings due to lack of ground-truth data by i) comparing against population statistics and ii) investigating real-world campaign events and how users participate political discussion in different cities. We found that the correlation between the population ratios and the user ratios on Twitter aligns significantly. The ones that are over estimated the most tend to be more populous cities, which may be a sign of technological and economic superiority, as those cities are the 3 of the four most "competitive" cities in Türkiye [27].

## References

[1] S. Stier, A. Bleier, H. Lietz, and M. Strohmaier, "Election campaigning on social media: Politicians, audiences, and the mediation of political communication on facebook and twitter," in *Studying Politics Across Media*. Routledge, 2020, pp. 50–74.

[2] F. Gilardi, T. Gessler, M. Kubli, and S. Müller, "Social media and political agenda setting," *Political Communication*, vol. 39, no. 1, pp. 39–60, 2022.

[3] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proc. of the Intl. Conf. on World Wide Web*, 2010, pp. 851–860.

[4] E. Ferrara, O. Varol, F. Menczer, and A. Flammini, "Traveling trends: Social butterflies or frequent fliers?" in *Proceedings of the first ACM conference on Online social networks*, 2013, pp. 213–222.

[5] Z. Tufekci, *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press, 2017.

[6] R. J. Gallagher, A. J. Reagan, C. M. Danforth, and P. S. Dodds, "Divergent discourse between protests and counter-protests:# blacklivesmatter and# alllivesmatter," *PloS One*, vol. 13, no. 4, p. e0195644, 2018.

[7] A. Karami, V. Shah, R. Vaezi, and A. Bansal, "Twitter speaks: A case of national disaster situational awareness," *Journal of Information Science*, vol. 46, no. 3, pp. 313–324, 2020.

[8] A. Culotta, N. Kumar, and J. Cutler, "Predicting the demographics of twitter users from website traffic data," in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 29, no. 1, 2015.

[9] E. Colleoni, A. Rozza, and A. Arvidsson, "Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data," *Journal of Communication*, vol. 64, no. 2, pp. 317–332, 2014.

[10] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[11] R. Fan, O. Varol, A. Varamesh, A. Barron, I. A. van de Leemput, M. Scheffer, and J. Bollen, "The minute-scale dynamics of online emotions reveal the effects of affect labeling," *Nature Human Behaviour*, vol. 3, no. 1, pp. 92–100, 2019.

[12] Q. Yuan, G. Cong, K. Zhao, Z. Ma, and A. Sun, "Who, where, when, and what: A nonparametric bayesian approach to context-aware recommendation and search for twitter users," *ACM Transactions on Information Systems*, vol. 33, no. 1, pp. 1–33, 2015.

[13] A. Anagnostopoulos, F. Petroni, and M. Sorella, "Targeted interest-driven advertising in cities using twitter," *Data Mining and Knowledge Discovery*, vol. 32, no. 3, pp. 737–763, 2018.

[14] O. Almatrafi, S. Parack, and B. Chavan, "Application of location-based sentiment analysis using twitter for identifying trends towards indian general elections 2014," in *Proc. of the Intl. Conf. on Ubiquitous Information Management and Communication*, 2015, pp. 1–5.

[15] Z. Gong, T. Cai, J.-C. Thill, S. Hale, and M. Graham, "Measuring relative opinion from location-based social media: A case study of the 2016 us presidential election," *PloS One*, vol. 15, no. 5, 2020.

[16] O. Varol, "Who follows turkish presidential candidates in 2023 elections?" in *2023 31st Signal Processing and Communications Applications Conference (SIU)*, 2023, pp. 1–4.

[17] H. Wei, J. Sankaranarayanan, and H. Samet, "Finding and tracking local twitter users for news detection," in *Proc. of the Intl. Conf. on Advances in Geographic Information Systems*, 2017, pp. 1–4.

[18] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in *Proc. of the ACM Intl. Conf. on Information and Knowledge Management*, 2010, pp. 759–768.

[19] S. Demirci and S. Ö. Özdemir, "An intelligent system for predicting location from text content on social media," in *Intl. Conf. on Computer Science and Engineering*. IEEE, 2017, pp. 671–676.

[20] O. Varol, E. Ferrara, C. L. Ogan, F. Menczer, and A. Flammini, "Evolution of online user behavior during a social upheaval," in *Proc. of the ACM Conf. on Web Science*, 2014, pp. 81–90.

[21] C. Budak and D. J. Watts, "Dissecting the spirit of gezi: Influence vs. selection in the occupy gezi movement," *Sociological Science*, vol. 2, pp. 370–397, 2015.

[22] O. C. Seckin, A. Atalay, E. Otenen, U. Duygu, and O. Varol, "Mechanisms driving online vaccine debate during the covid-19 pandemic," *Social Media + Society*, vol. 10, no. 1, p. 20563051241229657, 2024.

[23] O. Bas, C. L. Ogan, and O. Varol, "The role of legacy media and social media in increasing public engagement about violence against women in turkey," *Social Media + Society*, vol. 8, no. 4, p. 20563051221138939, 2022.

[24] A. Najafi, N. Mugurtay, Y. Zouzou, E. Demirci, S. Demirkiran, H. A. Karadeniz, and O. Varol, "First public dataset to study 2023 turkish general election," *Scientific Reports*, vol. 14, no. 8794, 2024.

[25] "Genel sorgu," https://postakodu.ptt.gov.tr/, accessed: 2023-12-31.

[26] "Adrese dayalı nüfus kayıt sistemi sonuçları, 2022," https://data.tuik.gov.tr/Bulten/Index?p=49685, accessed: 2024-01-22.

[27] A. N. Albayrak and G. Erkut, "Türkiye'de il ve bölgelerin rekabet gücü analizi," *İTÜDERGİSİ/a*, vol. 9, no. 2, 2011.