CENTRAL ASIAN JOURNAL OF MATHEMATICAL THEORY AND COMPUTER SCIENCES
https://cajmtcs.casjournal.org/index.php/CAJMTCS
Volume: 07 Issue: 01 | January 2026 ISSN: 2660-5309

Article

Big Data Analysis to Inferring Nationality Using X Social Network without GPS

Tareq Abed Mohammed[*1]

1. University of Kirkuk, College of Veterinary Medicine, Kirkuk, Iraq
* Correspondence: Tareq.mohammed@uokirkuk.edu.iq

Abstract: Inferring user's nationality from social media destinations gets to be a hot inquire about topic. In this paper we propose a modern and basic data analysis and algorithm to predict the nationality of X social network client without utilizing any GPS data like past proposed algorithms. The proposed algorithm employs the X social network user friends location data as it were. In spite of the fact that as it were around 30% of the X clients compose their location data in important form, we demonstrate that this percent is sufficient to de-cide the root nation or the nationality of any X client. Our proposed algorithm classifies more than 90% of the X client in our collecting dataset. We utilize within the proposed algorithm six nations but this work can effectively be generalized to incorporate all the world nations.

Keywords: Big Data, Database, Data Analysis, Machine Learning, KNIME, Social Media Analyzing

Citation: Mohammed T. A. Big Data Analysis to Inferring Nationality Using X Social Network without GPS. Central Asian Journal of Mathematical Theory and Computer Sciences 2026, 7(1), 173-182.

Received: 15th Nov 2025
Revised: 30th Nov 2025
Accepted: 12th Dec 2025
Published: 25th Dec 2025

1. Introduction

In today's interconnected world, social media platforms like X platform offer a treasure trove of data about users [1]. But what if you could use this data to understand something as personal as nationality, without relying on GPS information? This is where the power of big data inference comes in [2], [3].

This paper explores the fascinating world of inferring user nationality on X, solely through the analysis of big data. We'll delve into the various techniques employed, such as analyzing language patterns, time zone activity, followers' friends' information and even following trends and hashtags [4]. By harnessing the col-lective power of this data, we can build models that predict a user's nationality with surprising accuracy.

This research has significant implications for various fields. Imagine targeted marketing campaigns that resonate with specific nationalities, or even the ability to track international sentiment on a particular topic [5]. We'll explore the ethical considerations of such practices, ensuring responsible use of this powerful tech-nology.

In this paper we propose a modern calculation to decide the X client beginning nation area or we will say the nationality of the client depending on the companions (the X clients that this client takes after) of this client.

Related Works

These related works provide a broader understanding of the methods, techniques, and ethical considera-tions involved in inferring nationality from big data, complementing

the approach proposed in "Inferring Na-tionality on X: Big Data to the Rescue (without GPS)

Xilei Zhao et al in 2022. This study proposes an unused technique to analyze rapidly spreading fire departure utilizing large-scale GPS information. It categorizes evacuees and can illuminate crisis supervisors for superior readiness GPS information can give important bits of knowledge into departure behavior amid rapidly spreading fires [6].

Xiaolei Huang et al. The consider presents a multilingual Twitter corpus with creator statistic properties for despise discourse detection. It points to assess reasonableness in archive classifiers, considering age, na-tion, sex, and race. The dataset consolidates client profiles for gathering statistic variables, encouraging fair-minded classifier advancement. The think about investigates language varieties and statistic consistency, gives corpus insights, and surveys deduction precision for age, race, and sexual orientation. The discoveries lay the establishment for assessing reasonableness in hate discourse acknowledgment over dialects [7].

Joran Cornelisse and Raoul Grasman in 2020. The study aims to infer neuroticism of Twitter users based on their following interests. The researchers tested the hypothesis that following behavior on social media can reveal personality aspects. They used a two-step approach to collect data and build a regression model. The model was validated on a sample of Twitter users who filled out a personality questionnaire. Specific interests and occupations were used as predictors for neuroticism. Clustering techniques were used to simpli-fy the data. The study confirmed a correlation between following behavior on Twitter and neuroticism per-sonality dimension [8].

Kim, J., Sîrbu, A., Rossetti, G., & Giannotti, F, This paper analyzes the characteristics and behav-iors of vagrants and locals on Twitter, filling a hole in past investigate. The ponder incorporates a common evaluation of highlights such as profiles and tweets, as well as a broad organize examination. Discoveries incorporate that vagrant have more supporters, tweet more, and tend to associate based on nationality in-stead of nation of home. The investigate dataset utilized 200,354 clients from Twitter, with 4,940 distinguished as vagrants and 46,948 as locals. The ponder gives bits of knowledge into how these communities utilize Twitter and associated on social systems [9].

Mubarak et al in 2022, The think about centers on sex investigation and deduction on Arabic Twitter. The analysts analyze contrasts between male and female clients in terms of client engagement, subjects of in-trigued, and callings. They propose a strategy to gather sex utilizing usernames, profile pictures, tweets, and friends' systems. They physically commented on 166K Twitter accounts related with 92K client areas and accomplished an F1 score of 82.1% in sexual orientation deduction. The inquire about points to address the crevice in sex examination for Arabic Twitter and gives a modern dataset for assist ponder. The ponder high-lights the significance of statistic data in decision-making forms and the require for programmed strategies for sexual orientation deduction on social media. The analysts compare their work to past ponders and em-phasize the uniqueness of their dataset and examination strategies in sexual orientation deduction for Arabic [10].

In [11] Inyoung Jun et al in 2023, The article talks about assessing open discernments of pesticide utilize, security, and control on Twitter. It analyzes tweets related to pesticides between 2013 and 2021, categoriz-ing them into distinctive client bunches and organizations. The consider points to get it communication be-haviors, assumptions, and dialog points related to pesticides on social media. Open discernments are general-ly negative, with person accounts centering on wellbeing and natural dangers, whereas industry and gov-ernment accounts emphasize agrarian utilization and directions. The ponder highlights the significance of understanding open opinions, needs, and recognitions to progress communication and decision-making with respect to pesticides.

Rochana Chaturvedi and Sugat Chaturvedi in 2024 [12], The ponder employments machine-learning models to gather religion from names in South Asia, permitting for

classification of concealed names. The phonetic roots of names are analyzed to get it the designs related with diverse religions. The approach is connected to Indian discretionary candidate names, uncovering a decrease in Muslim representation. Names are imperative markers of gather personality and can be utilized to think about devout demography and sep-aration in welfare program assignment. The models beat existing strategies and can be connected to gather other markers of bunch character. The ponder moreover highlights the significance of religion in forming in-clinations, demeanors, and results.

2. Methodology

Dataset Collection

One of the preeminent basic and troublesome steps is collecting the dataset. As of late, X has controlled anyone from open sharing of the X client data, undoubtedly, in case that data is for exploration, which makes it uncommonly troublesome to find the X users' database that contains information of various X clients [13]. This allows any examiner to download the data by themselves from the X API, which takes a long time after the confinements were included by X, such as the square API for 15 minutes after 15 client friends' requests [14].

In this paper, a magnificent data examination tool was utilized, named KNIME [15]. KNIME is a graphical client interface license that gets together nodes for data preprocessing. It can be utilized for modeling and visualization [16].

For planning, open information of 250,000 X clients was collected utilizing unused KNIME X nodes, which unravels the download plan. The collected dataset joins a 1500 X client with their Followers and companions. We select the clients erratically by searching for the word inside the X utilizing the X node, which calls the API Xs. After that, we filter by erasing any X client on the off chance that they did not provide their location. We keep our work in this study because it was on six countries, which are the United States, Canada, England, Australia, Japan, and India. Figure 1 shows the rate of clients from each country.

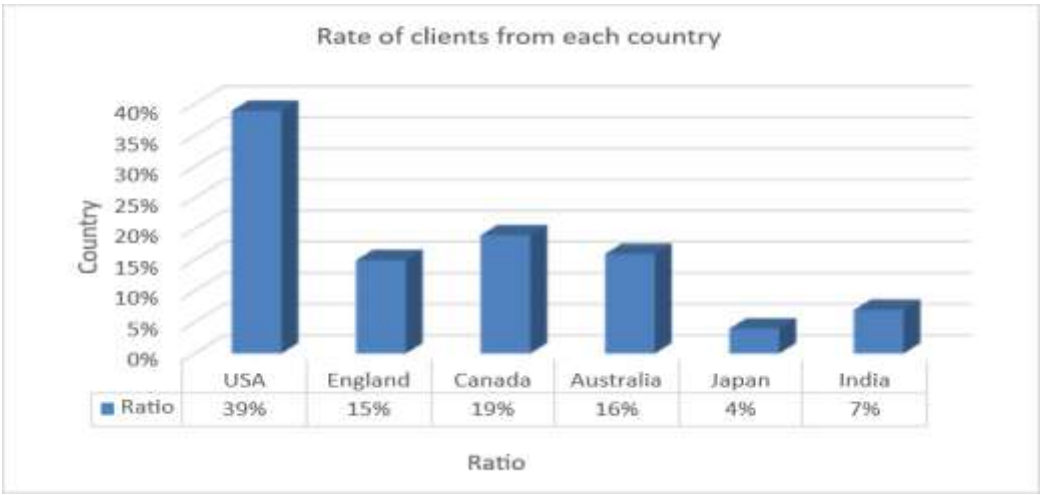

Figure 1. X user distribution over countries.

To collect the friend's information of each X client, R scrap was utilizing a node from KNIME [16-20]. This node grants execution of an R script from the interior of KNIME. We compose R code to inquire the X client friends and supporters utilizing X(R) [17-19-21] library to call X API functions. To test the proposed calculation, another set of 120,000 X client information is collected with the same steps.

3. Results and Discussions

Experiments and Analysis

When a course of action chooses Client X's location or nationality, we are going clarify that Client X's na-tionality can be expelled from its friend's location and preempted. A huge issue with this approach is that a few X clients do not indicate their location in their account at all, and a few of them enter an outside loca-tion. In our dataset, we found that 30% of X clients don't sort area data, and a few sorts futile words rather than areas. Within the information set we collected, roughly 35% of X clients shaped key positions in their accounts.

In truth, when clients type in their areas, they type in them in a totally diverse fashion since laying out cli-ents can compose the title of the country, the capital of the country, the title of the town, at most for these choices A combination. So, extricating country names from X users' areas isn't an essential assignment.

In our examination, we began with utilize the collected arranging information (as portrayed within the past section) to decide the proportion of companion locations that decides the nationality of X client. To do this we take after these steps:

1. We partition clients in each country into a few bunches, check and analyze the area data composition of X clients in each country.
2. We delete the X clients that have exceptionally huge number of devotees or companions since as a run the work, the companies and educate will have this gigantic number of supporters. And presently and after that within the occasion that that X client proprietor may be a person and not an organized, he will be a celebrated person and customarily his location or nationality known.
3. We delete X clients having a horrendously small account of friends since this small number isn't sat-isfactory to choose the client location, in this way may causes a desire mistake.
4. For each chosen country, we select many words fitting for that country mold to appear that this composed location implies the found country. Table 1 shows up outlines of the utilized words for each country. We for the foremost portion utilized the country full title in English, country brief title, country full title in that country tongue, the gigantic cities names and a number of celebrated places. Concurring to these key words for each country, the devotees and friends' zones of each X client are chosen. After that we collect this information and put it in six set freely one set for each country.

Tabel 1. Searching key words used to classify the X users' countries.

Country	Examples of Keywords
USA	Miami – Washinton DC –Texas
England	Manchester - London
Canada	Ottawa – Toronto
Australia	Sydney
Japan	Tokyo
India	Mumbai - Delhi- Abad- Pune – Kerala

In organize to depict the ultimate step in inconspicuous components, (F_1, F_2 ... F_N) Mean the companions of a X client (U) and let us classify each companion Fi of this X client (U) concurring to the recorded word keys inside the previous table. In case companion (F_i) composes on his X area one of the words appearing the country of client (U), at that point ($F_i = 1$) else (F_i will be zero). After classifying all the companions (F_i) of the client (U) we

compute the taking after regard (Uloc) to illustrate how various friends appear to sort within the same area of the client U.

$$U_{Loc} = \frac{1}{N}\sum_{k=0}^{N} F_i \; ; \; F_i \in [0,1] \qquad (1)$$

This value U_{Loc} represents how the friends of a twitter user may be able to indicate the country of that twitter user. Figure 2 shows the histograms of the resulted twitter user classifying according to this value. For example, most of USA twitter users get a U_{Loc} value between 10% and 30% as the big two middle columns show, where the number of users that values from 30 to 40 is small, and the number of users that get from 40 to 50 is very small. We can also note the huge difference between countries which reflects their user habits in writing their locations.

This regard U(Loc) talks to how the companions of a X client may be able to appear the country of that X client. Figures (2 – 7) shows up the histograms of the brought almost X client classifying concurring to this regard. We are ready in addition note differentiate between countries that reflects their client affinities in composing their countries.

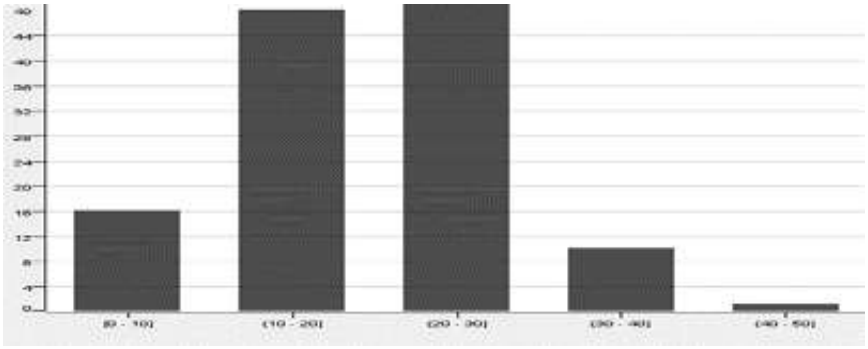

Figure 2. Percentage of X user friend's histograms after classifying for USA.

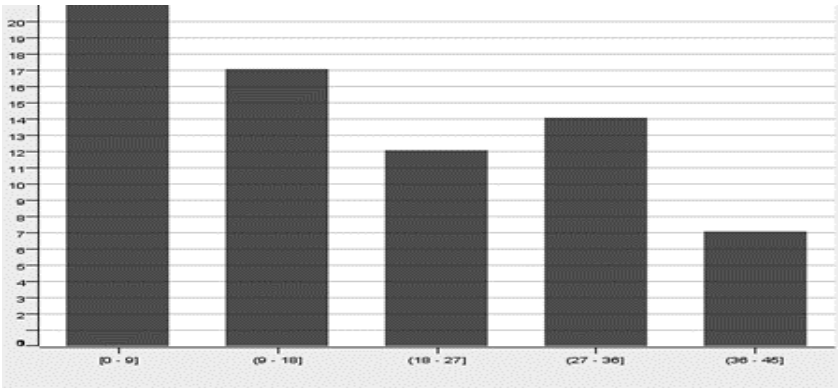

Figure 3. Percentage of X user friend's histograms after classifying for CANADA.

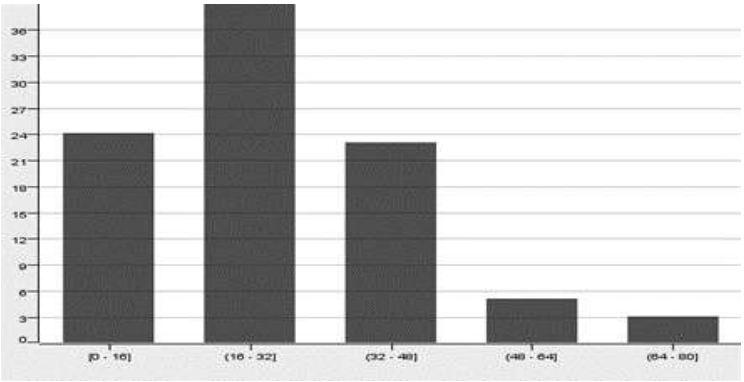

Figure 4. Percentage of X user friend's histograms after classifying for ENGILAND.

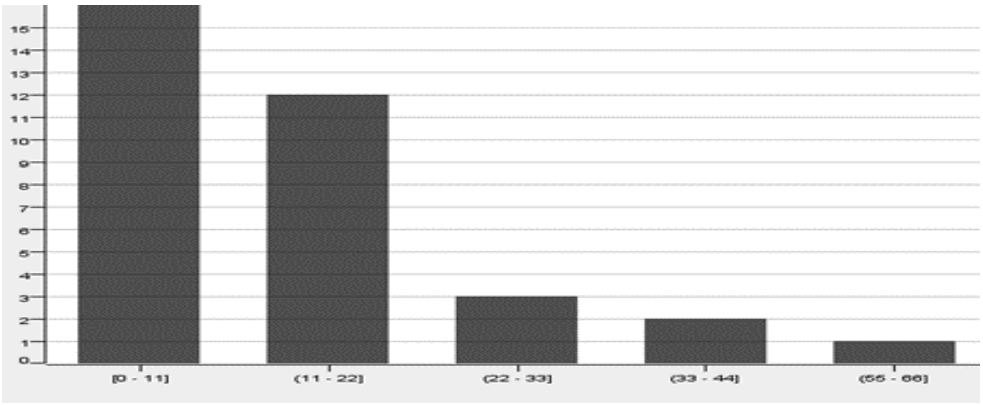

Figure 5. Percentage of X user friend's histograms after classifying for AUSTRALIA.

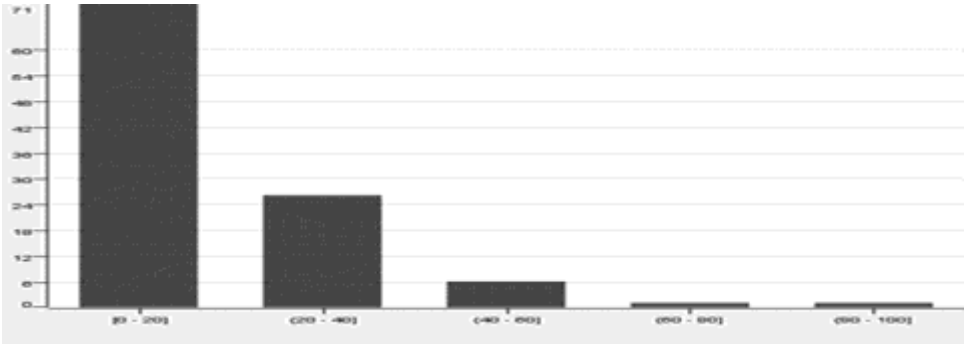

Figure 6. Percentage of X user friend's histograms after classifying for INDIA.

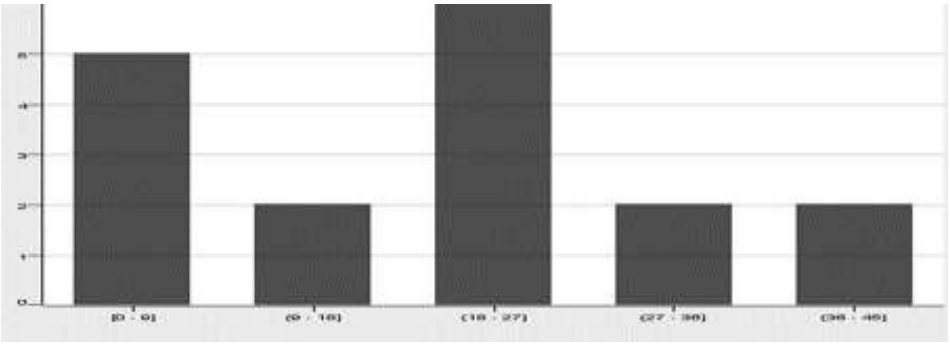

Figure 7. Percentage of X user friends' histograms after classifying for JAPAN.

Figures (8-13) show the box plot of the U(Loc) values for each country utilizing the interquartile range (IQR). As well as the tremendous qualification between countries, in addition being clear. We note for the case that the finest country is Australia (box edges between 15 and 40) since most of its people sort in their country title explicitly and in a horrendously comparable mold. Where in other countries like India, the number of clients who sort in critical areas is small (box edge between 3 and 20). Concurring with this result is troublesome to find an edge to the regard U(Loc) to classify X clients from all countries. So, we make a cutting-edge methodology to choose the area of an X client according to the U(Loc) regard of that client for each country, as depicted in taking after calculation.

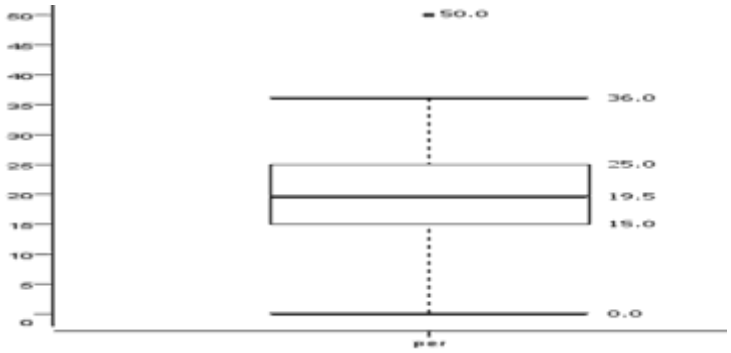

Figure 8. The box plot of the U_{Loc} values for the USA.

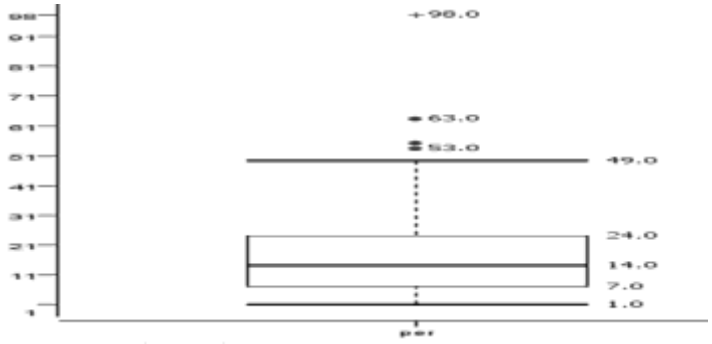

Figure 9. The box plot of the U_{Loc} values for CANADA.

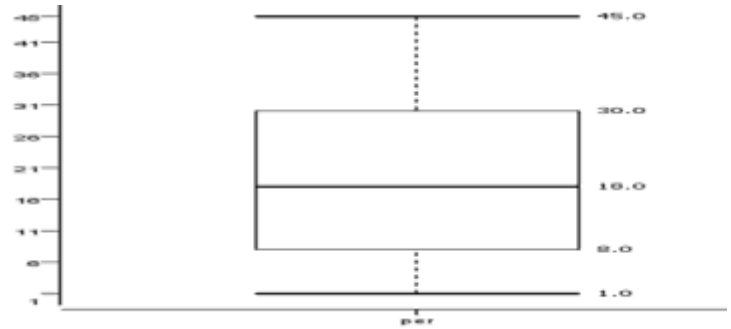

Figure 10. The box plot of the U_{Loc} values for ENGLAND.

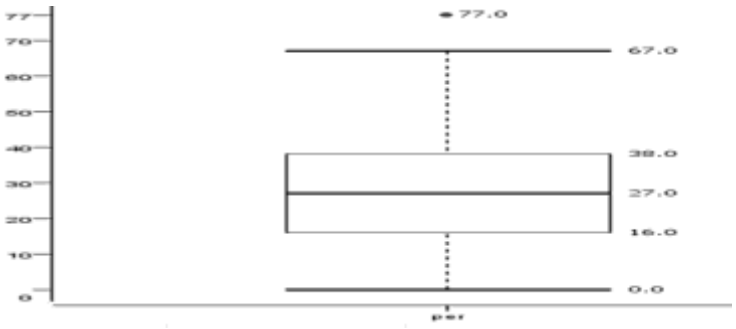

Figure 11. The box plot of the U_{Loc} values for AUSTRALIA.

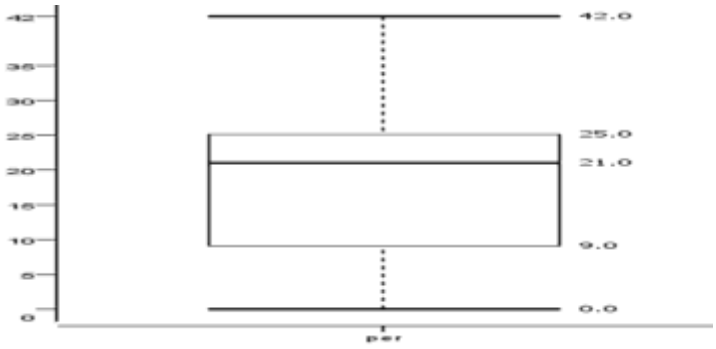

Figure 12. The box plot of the U_{Loc} values for INDIA.

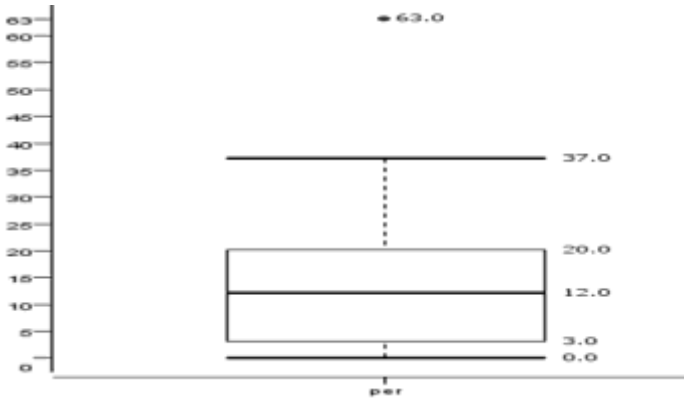

Figure 13. The box plot of the U_{Loc} values for JAPAN.

To classify the area of any X client U:

1. Get all the companions' areas of the client U.
2. Apply the looking calculation to the areas corresponding to the component, which are classified as of now depending on cities, capital, and country names, as shown in Table 1.
3. Find the regard U(Loc) that is defined in the Equation for each country.
4. If the foremost extraordinary U(Loc) is more unmistakable than the slightest regard V(min) of that country at then recognize this X client area as the country that has the most noteworthy U(Loc).
5. In case the foremost extraordinary U(Loc) is more diminutive than the slightest regard of that country V(min) at the point test, the following most raised U(Loc) regard nation, implied as U(Loc2). On the occasion that U(Loc2) is more critical than their country V(min), recognize this X client area as the country that has the taking after most noteworthy U(Loc2).
6. Repeat steps (4 and 5) until you circle the countries for client U.

The slightest values V(min) to the U(Loc) can be chosen in a few steps. In this ponder, V(min) values were chosen to rise to the lower edge of the box plot. For the outline, for the USA, the V(min) = 16% whereas for India, V(min) = 5%. Let us give two outlines, within the occasion that we find the U(Loc) values for a cer-tain client and found as (U(Loc)=12% for USA, U(Loc)=11% for Britain, U(Loc)=7% for Canada) at that point we'll classify this X client area to be Britain since V(min) for USA = 15% and so we go to the minute most essential country which is Britain. For Britain, V(min) = 7% which is less than U(Loc) for that country. Within the minute outline, within the occasion that we have these values for another X client (U(Loc)=19% for USA, U(Loc)=11% for Britain, U(Loc)=7% for Canada) we'll obviously select USA as the area of this client since it has the most noteworthy U(Loc) regard conjointly more critical the V(min) of USA.

Inside the moment attempt, we utilize the moment collected dataset, which contains information of 120,000 X clients and their companions from the same over six recorded countries. After applying our calculation, we appear to precisely classify 92% of the X clients. in this paper, the exploration work is constrained to six countries for modifications, but this work can easily be expanded to include all countries.

4. Conclusion

A modern algorithm proposed from data analysis to recognize the nationality of a X social network client utilizing the locations of their companions. A modern algorithm is proposed to compute a regard for each candidate nationality for a certain X client. The nationality is chosen based on data analysis pre-calculated slightest values for each country. The proposed algorithm viably classifies 92% of the X clients inside the dataset. In this work we utilized the friends' locations and we'll incorporate some information to get more correct results almost as future work.

REFERENCES

[1] M. Aria, C. Cuccurullo, L. D'Aniello, M. Misuraca, and M. Spano, "Thematic analysis as a new culturomic tool: The social media coverage on COVID-19 pandemic in Italy," *Sustainability*, vol. 14, no. 6, art. no. 3643, 2022.

[2] C. Qian, N. Mathur, N. H. Zakaria, R. Arora, V. Gupta, and M. Ali, "Understanding public opinions on social media for financial sentiment analysis using AI-based techniques," *Information Processing & Management*, vol. 59, no. 6, art. no. 103098, 2022.

[3] M. T. Refsnes, *Whitman on TweetDeck: Community and Self Through Walt Whitman's Leaves of Grass and Blogging*, M.S. thesis, Univ. of Bergen, Bergen, Norway, 2022.

[4] V. Mathur, C. Lustig, and E. Kaziunas, "Disordering datasets: Sociotechnical misalignments in AI-mediated behavioral health," *Proc. ACM Hum.-Comput. Interact.*, vol. 6, no. CSCW2, pp. 1–33, 2022.

[5] C. Cross and M. Lee, "Exploring fear of crime for those targeted by romance fraud," *Victims & Offenders*, vol. 17, no. 5, pp. 735–755, 2022.

[6] X. Zhao *et al.*, "Estimating wildfire evacuation decision and departure timing using large-scale GPS data," *Transportation Research Part D: Transport and Environment*, vol. 107, art. no. 103277, 2022, doi: 10.1016/j.trd.2022.103277.

[7] T. A. Mohammed, S. Alhayli, S. Albawi, and A. D. Duru, "Intelligent database interface techniques using semantic coordination," in *Proc. 1st Int. Scientific Conf. Engineering Sciences–3rd Scientific Conf. Engineering Science (ISCES)*, Jan. 2018, pp. 13–17.

[8] X. Huang, L. Xing, F. Dernoncourt, and M. J. Paul, "Multilingual Twitter corpus and baselines for evaluating demographic bias in hate speech recognition," *arXiv preprint*, arXiv:2002.10361, 2020.

[9] J. Cornelisse, "Inferring neuroticism of Twitter users by utilizing their following interests," in *Proc. 3rd Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, Dec. 2020, pp. 1–10.

[10] J. Kim, A. Sîrbu, G. Rossetti, and F. Giannotti, "Characterising different communities of Twitter users: Migrants and natives," in *Complex Networks & Their Applications X*, Cham, Switzerland: Springer, 2022, pp. 130–141.

[11] H. Mubarak, S. A. Chowdhury, and F. Alam, "Arabgend: Gender analysis and inference on Arabic Twitter," *arXiv preprint*, arXiv:2203.00271, 2022.

[12] I. Jun *et al.*, "Evaluating the perceptions of pesticide use, safety, and regulation and identifying common pesticide-related topics on Twitter," *Integrated Environmental Assessment and Management*, vol. 19, no. 6, pp. 1581–1599, 2023.

[13] H. Al-Bayaty, T. Mohammed, A. Ghareeb, and W. Wang, "City-scale energy demand forecasting using machine learning-based models: A comparative study," in *Proc. 2nd Int. Conf. Data Science, E-Learning and Information Systems*, Dec. 2019, pp. 1–9.

[14] R. Chaturvedi and S. Chaturvedi, "It's all in the name: A character-based approach to infer religion," *Political Analysis*, vol. 32, no. 1, pp. 34–49, 2024.

[15] P. Vyas, G. Vyas, and G. Dhiman, "Ruemo—The classification framework for Russia–Ukraine war-related societal emotions on Twitter through machine learning," *Algorithms*, vol. 16, no. 2, art. no. 69, 2023.
[16] "Early prediction of stroke risk using machine learning approaches and imbalanced data," *NTU-JET*, vol. 4, no. 1, Mar. 2025, doi: 10.56286/1vf19469.
[17] A. K. Abbas, T. A. Mohammed, O. Bayat, and O. N. Ucan, "The prediction of fusion degree of international groups from their Twitter accounts," in *Proc. Int. Conf. Engineering and Technology (ICET)*, Antalya, Turkey, 2017, pp. 1–8, doi: 10.1109/ICEngTechnol.2017.8308185.
[18] KNIME Analytics Platform, "KNIME," [Online]. Available: https://www.knime.com. Accessed: Feb. 20, 2024.
[19] B. Maraza-Quispe *et al.*, "A predictive model implemented in KNIME based on learning analytics for timely decision making in virtual learning environments," *Int. J. Information and Education Technology*, vol. 12, no. 2, pp. 91–99, 2022.
[20] K. K. Abbo and H. H. Mohamed, "New scaled conjugate gradient algorithm for training artificial neural networks based on pure conjugacy condition," *Kirkuk Journal of Science*, vol. 10, no. 3, pp. 230–241, 2015.
[21] E. Z. Mohammed, "Proposed classification system by using artificial neural network," *Kirkuk Journal of Science*, vol. 10, no. 3, pp. 59–78, 2015.