**Twitter Data Project**

The Jordan Center for the Advanced Study of Russia was established in 2011 thanks to a generous gift from the family of NYU alumni Boris and Elizabeth Jordan. The mission of the Center is to make Russia intrinsic to all aspects of scholarly investigation: from history to visual culture, literature to economics, anthropology to politics. The Center also engages with scholars and projects that examine critical questions pertinent to the broader post-Soviet region.

One such question is what role social media plays in political outcomes, including election results and discourses around candidates, parties, and elections. Availability of technology and advances in machine learning make answering these questions possible for countries and regions around the world, including the post-Soviet region. Comparison of the role and patterns of social media use in politics across various countries in the post-Soviet space can provide a unique understanding of the political dynamics for scholars and practitioners alike. However, before proper comparison can be made across countries, we must understand the dynamics within each country. One such country of interest for cross-national comparison is Mongolia. This project makes the broad question of the political role of social media tractable by exploring how election-related discourse unfolded on Twitter in Mongolia around the parliamentary election of 2020.

Mongolia has been called an oasis of democracy for its unique position as a democratic country, despite being surrounded by authoritarian regimes in its immediate and near neighborhood. It has successfully held legislative elections since 1992 every four years. The parliamentary election of 2020 was the eighth democratic election that the country held to determine who would lead the country. Mongolia has two main parties—the Mongolian People’s Party (MPP) and the Democratic Party (DP). The former is a remnant of the Communist Party, while the latter emerged as an opposition to support democratic model of governance in the early 1990s when the Soviet Union disintegrated. A third political party has emerged in recent years—Zuv Khun (ZKH) Electorate. Overall, 606 candidates registered with the General Election Committee in the 2020 election, 121 of which were independents, and the rest ran from 13 political parties and four coalitions. The election was organized following a plurality system with 29 multi-member districts. The MPP cemented its position as a dominant party in Mongolian politics by winning 62 out of 76 seats. The DP received 11 seats. One candidate from Zuv Khun qualified to occupy a seat. The remaining two seats went to an independent and a candidate from a fourth minor party.

**Dataset**:

Twitter data were scraped during the election of 2020 with following keywords: the names of the three main parties (DP, MPP, and ZKH) and “#Сонгууль2020,” which means #Election2020 in Mongolian language.

The collected twitter data is mostly in Mongolian, but the data mining tool captured some content in Russian and other languages. Those tweets in different languages should be separated from the dataset. For this purpose, we encourage the team to explore new and emerging machine learning capabilities to detect various languages in order to remove the tweets in languages other than Mongolian.

We do not have formal data dictionaries of this data. However, most of the variables should be based on twitter meta-tags. See Murphy (2016) for coding ideas.

**Questions for students:**

We are interested in anything of interest you find in the data quantitatively. Explore the options to apply various machine learning models on the dataset for any interesting and useful insights. See Lassen et al. (2016).

Items of particular interest for our research, include the responses to some of the following questions:

* Who participated in the twitter discussions? Specifically, who are the most frequent Twitterers? How many followers did these top twitterers’ discourse reach?
* Daily tweet counts following election day (What tweets were the most popular for each of the three hashtags for the MPP, DP, and ZKH? Did the intensity of discourses decline as the days passed after election day?)
* The average length of the tweets each day (Are people more or less interested in discourses on these political parties as days pass after election day?)
* Whose tweets were tweeted the most? (based on user\_followers\_count, user\_friends\_count and see other columns etc.,)
* Do we observe variation in the geographic location of the tweets across the three groups? (e.g. people abroad may move on to other topics soon after election vs. people on the ground in Mongolia)
* How many unique tweets, re-tweets?
* How many of them had the link to news, images, videos?
* How many different hashtags? Cluster by hashtags?
* How many of them used hashtags?
* Clustering by the twitter handle of each party and possibly by their slogans.

We are ready to answer any questions about the data and analysis strategies. We would like to have weekly updates from the team and schedule the virtual meetings at the critical points of your analysis. At the end, we look forward to a presentation.

**Expected deliverables**:

* Clean dataset for further textual analysis
* Propose unique ML based analysis to Dr. Uvsh for approval
* Interim reporting mid-semester
* Reportable/publishable analysis of data
* Presentation of final analysis
* Final report

**Contact person**:

Dr. Delgerjargal Uvsh

Postdoctoral fellow

Jordan Center for the Advanced Study of Russia

New York University

Email: [du2026@nyu.edu](mailto:du2026@nyu.edu) (this email will work until the end of October), [uvsh@wisc.edu](mailto:uvsh@wisc.edu) (This is my permanent email.)

Phone: (812)-606-4077

**Suggested reading**:

Election 2020 in Mongolia:

<https://www.washingtonpost.com/politics/2020/07/07/heres-how-an-unpopular-ruling-party-swept-mongolias-june-elections/>

Excellent reading on social media and democracy in general:

<https://www.google.com/books/edition/Social_Media_and_Democracy/NEjzDwAAQBAJ?hl=en&gbpv=1&printsec=frontcover>

Lassen, N., Cour, L. & Vatrapu, R. (2016). Predictive analytics with social media data. In *The SAGE Handbook of social media research methods* (pp. 328-341).

Murthy, D. (2016). The ontology of tweets: mixed methods approaches to the study of twitter. In *The SAGE Handbook of social media research methods* (pp. 559-572).

**Fairness/Bias Assignment**

**Data**

1. For each Twitter user in the dataset extract the following additional information through the Twitter API: year joined and user location.
2. Create 2 new features and add to the datasets: number of years using Twitter (**numyears**) and the user location in their Twitter profile (**loc**).  
   Note: there is also a “user location” column in the data but that appears to be their current location and not their home location as per the profile.
3. Analyze the data with respect to the **loc** feature for the Twitter users and compare with the population (Mongolia) distribution by location. Use publicly available country data for the comparison.
4. Consider what the **numyears** feature can signify – age, experience with using Twitter, etc.
5. Focus on the new features **numyear**s and **loc** as the sensitive features.
6. Report your results with visualizations that highlight the key differences between the sample and general populations for geographic distribution.
7. Based on your results, is there bias in the sample data?
8. Analyze the data by a combination (2) of features (sensitive and other). Often features that are not considered sensitive can be a proxy for sensitive features. For example, a difference between loc and user location (travel) could indicate a higher income or more educated user.
9. Determine feature groupings that are relevant for your project and explain your choices.
10. Can you derive any bias analysis features from the Twitter hashtags?
11. Based on this data analysis, develop a hypothesis about where fairness/bias issues could arise in the models you have in the primary assignment.

**Modeling**

1. Based on your hypothesis, assess the fairness of your models by applying the fairness-related metrics that are available in the Python **Fairlearn** package and/or the R **Fairness** package or with other similar tools.
2. Explain the reasoning for the groups that you selected for the fairness metrics.
3. Compare the fairness metrics for the different groups.
4. If you developed multiple models compare the fairness metrics for the models.
5. Comment on the results.
6. Suggest how the bias/fairness issues could be mitigated.
7. Present the results visually to show salient insights.