ISSUES IN IDENTIFYING THE AGE OF TWITTER USERS IN UKRAINE

Bogdan Pavliy, Toyama University of International Studies
Jonathan Lewis, Hitotsubashi University

Abstract

This article presents the results of the second stage of an ongoing project to examine language use by social media users in Ukraine. We use data from the microblogging service Twitter to analyze the age and language preferences of Twitter users in Ukraine. Based on a dataset of more than million geotagged tweets collected from April to September 2015, we studied the relation between the tweeting activity and the examinations in high-schools and universities. We found that the tweeting activity is related to the periods of external independent testing (in Ukrainian ЗНО) rather than to high-school and university examinations, but could not find the evidence that a considerable portion of users are high school senior students, or graduates.

Introduction

Widespread use of social media including Twitter in Ukraine had great social and political significance during the events on Euromaidan and other protesters meetings. The research on the use of social online communities such as Twitter or Facebook (Bachmann and Lyubashenko, 2014; Goban-Klas, 2014; Lyubashenko in Valenzuela et al., 2014; Ronzhyn, 2014, 2016) along with the Yandex report on the Twitter usage and trendiest Twitter hashtags in Ukraine in 2013-2014, is focused on political or socioeconomic aspects of use rather than gender or demography of users. Gender, geographic and demographic aspects of social media users in Ukraine are considered in the research of Onuch on the characteristics of Euromaidan protestors and ways of mobilizing them through social online communities (Onuch, 2014, 2015, 2015). However, in this type of research, both linguistic and demographic aspects are limited, as those involved in online communications – whoever they are: organizers, participants or supporters are on one side of the barricades in their political preferences, and cannot represent all nation in general. As the demographic aspect of Twitter users cannot be limited to those who protested against Yanukovych government during the Revolution of Dignity, is still important to investigate on a national level who is using Twitter in Ukraine and what are their basic demographic characteristics.

Concerning the research on Twitter use in Ukraine on the national level, the most interesting one was conducted by Pentina, Basmanova and Zhang (2014). In their cross-national study of Twitter users' motivations and continuance intentions the researches compare motivations and preferences of Twitter users in Ukraine and the United States. The authors did not opt for geotagged tweets but provided online surveys for selected Twitter users in both countries. Based on the previous research of Kostenko (2011), Pentina, Basmanova and Zhang (2014) suggest that "while typical Ukrainian Twitter user demographics are not available, the sample characteristics are representative of the Ukrainian Internet user who is characterized by a younger age and higher socioeconomic status that provides access to wireless and mobile communications" (Kostenko, 2011; Pentina, Basmanova and Zhang, 2014). However, their study does not directly relate to our research and still gives us no detailed understanding of the demographic characteristics of Twitter users on the national level. This study aims to contribute to filling that gap.

In the first stage of our research, published in 2015, we used the location and language information of publically available geotagged Twitter posts and mapped the Ukrainian and Russian tweets sent in different parts of the country (Pavliy and Lewis, 2015). We discussed the advantages and limitations of using social media data to investigate communicative behavior and geography of language use in Ukraine and came to conclusions that with the limitations of our data, we cannot claim that tweeting activities can be representative of the country's population as a whole (Pavliy and Lewis, 2015). Because it is possible that our sample contains a higher proportion of Russian native speakers to the extent that urban areas in Ukraine have a higher proportion of Russian native speakers than rural areas.

Purpose and aims of research

In this article we do not tackle any identity issues of the Ukrainian population but target to find some demographic information available about Twitter users and undertake the research on rather national than local level. We cannot clarify if the Twitter users, who send geotagged tweets, can be considered a representative sample of Ukrainian society. As we can see from the previous research, the lack of gender, social status and demographic data is a serious hindrance to integration of social media data into the social sciences (Sloan and Morgan, 2015). We still have to deal with the same questions as the first researchers on social media dealt with: "Are Twitter users a representative sample of society? If not, which demographics are over- or underrepresented in the Twitter population" (Mislove et al., 2011).

Establishing age and other demographic characteristics of Twitter users is not easy. A group of researchers from the UK developed "techniques for collecting or estimating demographics from

Twitter data including analyzing gender, language and location" (Sloan et al. 2013). Building on this work and adapting techniques used by other researchers, Sloan et al. (2015) developed a Twitter user age detection algorithm based on a set of pattern matching rules. They validated the reliability of their algorithm using expert human testers. Applying their technique to British Twitter users, they found that the age distribution of Twitter users is much younger than the UK population as of the 2011 Census, with a peak around ages 16 to 22 accounting for 67.5% of all users. They claimed that more than half of the users (59.4%) belonged to the age group 13 to 20. Their results also pointed to the existence of over half a million Twitter users over the age of 40 in the UK. However they conceded that due to the limitations of their detecting technique, the desirable accuracy can be attained only by analyzing the content of these tweets cross referencing with social survey data.

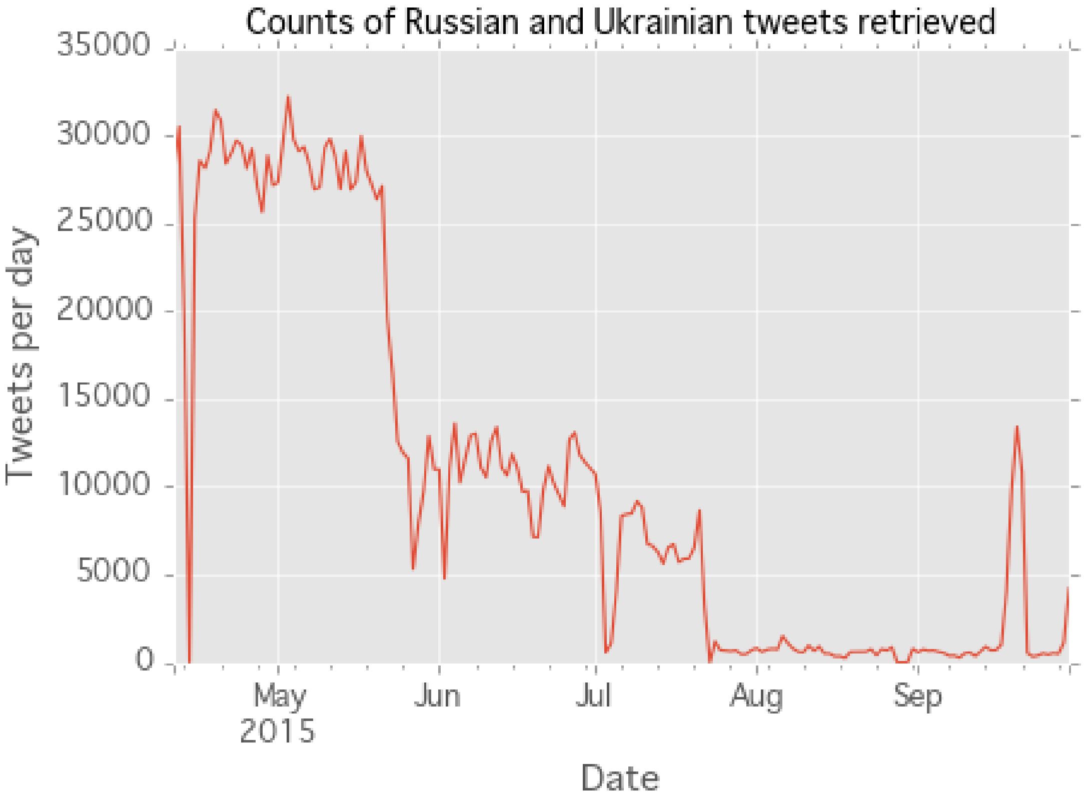
We understand that it is not a realistic task in the short term to reproduce their algorithm for Ukraine tweets as their techniques cannot be just imitated and applied to bilingual language environment in Ukraine. On this stage of our study we analyze geotagged tweets sent from the territory of Ukraine, considering them in both demographic and linguistic contexts. We also discuss the relationship between the examination periods and peaks of tweeting activity,　establishing the method of detecting the age of Twitter users through the use of words "exam" or External independent evaluation or External independent testing (in short "ЗНО" in both Ukrainian and Russian) in the content of geotagged tweets.

Procedure

We used Twitter's Streaming API to collect geotagged tweets sent from within the territory of Ukraine (including Crimea) between 1 April and 30 September 2015. We deleted some obviously "robot" tweets such as those sent by the FourSquare application. The resulting dataset comprised 2,458,953. Twitter tags each tweet with a language, the result of its own language detection algorithm; we have previously shown Twitter's language tags to be acceptably accurate with regard to Russian and Ukrainian tweets (Pavliy and Lewis, 2016). 1,553,787 or 63.2% of the collected tweets were in Russian and 242,829 or 9.9% were in Ukrainian; English tweets accounted for 8.7%, followed by a long tail of other languages such as Slovenian, Polish and Bulgarian with less than 3% each. 5.7% of the tweets were tagged as "language unidentified". The Ukrainian and Russian tweets were sent by a total 70,429 distinct users. As we mentioned in our previous research, when we consider that geotagged tweets account for only a few percent of all tweets sent, it is clear that we are observing the behavior of more than a tiny group of enthusiasts (Pavliy and Lewis, 2015).

The number of tweets retrieved each day shows great variation over the six-month period, as Figure 1 shows.

Figure 1 Counts of Russian and Ukrainian tweets



We see much higher activity between April and June compared to the summer months. While our data collection was not running during 100% of the period – as shown by the occasional dips of the graph to 0 – it was not the case that data collection was interrupted more frequently in the summer months than in the spring.

Although there are some regional differences in tweeting activity, the April-June peak is a nationwide phenomenon.

We hypothesized that the main reason for this seasonal variation in the number of tweets might be the periods of high school and university examinations conducted in May – June, and starting the new academic year in September. Ukrainian high school and university students have their examination period in May and June, so the reason for tweeting may be the queries about the content of the exams, requests for assistance, or, most probable, expressing to friends and classmates personal feelings and emotions related to the exams. The period of entrance examinations to the universities in 2015 was from July 10 to September 1, where the decrease in tweets is observed. However, the high school examinations and especially EIT (external independent testing) examination could possibly make a crucial impact on the increase of tweets in April-May. Subsequently, if our suggestion proves to be correct, we can state that a reasonable amount of geotagged tweets has been sent by high school students, graduates or university students during or before their examination period.

Examinations in Ukraine can be divided into six groups: high school examination, high school graduation examinations, external independent evaluation or external independent testing = EIT (in Ukrainian ЗНО), university entrance examinations, university examinations, and university graduation examinations. Among them the most intense ones are the EIT examinations, because they are unified for all educational facilities, they are independent, not biased or influenced by the personal relationship with the examiner and their results impact the future career of the examinee as they cannot apply to the universities if they have lower than threshold score. Table 1 shows the dates of the EIT "ЗНО" high school graduates examinations in Ukraine in 2015.

Table 1. ЗНО (EIT) Schedule in 2015

Discipline	Number of participants	Date of exam	Exam result announcement
Ukrainian language & literature (basic) (advanced)	267,394 21,583	24-04-2015	13-05-2015
French	840	03-06-2015	25-06-2015
German	3,172	05-06-2015	25-06-2015
Spanish	179	08-06-2015	25-06-2015
English	81,318	10-06-2015	25-06-2015
Math (basic) (advanced)	129,142 17,650	12-06-2015	25-06-2015
Russian	3,645	15-06-2015	03-07-2015
Biology	98,372	17-06-2015	03-07-2015
History of Ukraine	158,556	19-06-2015	03-07-2015
Physics	51,463	22-06-2015	03-07-2015
Geography	65,541	24-06-2015	08-07-2015
Chemistry	39,730	26-06-2015	09-07-2015

To check the correctness of our suggestion, we calculated how often words “exam” or “ЗНО” appear in the content of the tweets. Depending on the results, we can suggest what percent of geotagged twitter users are of high school/university age and to what extent the examination period impacted the tweeting activity in Ukrainian communities in 2015.

We found that the word “exam” “экзамен” (or its derivatives, in lower and upper case) in Russian appears in 6,375 tweets and “екзамен” (or its derivatives) in Ukrainian appears in 610 tweets. Moreover, in 90 tweets which were identified as Russian the word “екзамен” (or its derivatives) is spelled in the Ukrainian way. In those cases the senders presumably either did not know the correct spelling of the word exam in Russian, or wrote it deliberately in the Ukrainian manner.

We also counted occurrences of the usage of “ЗНО” – the Ukrainian/Russian abbreviation for external independent evaluation or external independent testing. Our count was case-insensitive and excluded instances where “зно” was used as an adverbial suffix. We found 2,931 occurrences in Russian and 1,328 in Ukrainian, which means that almost one third of all tweets with the word “ЗНО” were sent in Ukrainian.

Third, we calculated the number of tweets including words EXAM or its derivatives (“ЭКЗАМЕН” in Russian and “ЕКЗАМЕН” in Ukrainian) and “ЗНО” on weekly basis. We also checked if the increase in tweets is related to the most intense “ЗНО” exam periods.

Table 2. The results on tweets with the words EXAM and/or ЗНО

WEEK	Total Number of tweets	Total Number of tweets with EXAM + ЗНО	% of Total	Tweets with Exam in RU	Tweets with Exam in UK	Tweets with ЗНО	Tweets with ЗНО % of Total
2015/4/12	59947	69	0.12	36	1	32	0.05
2015/4/19	166799	634	0.38	230	31	373	0.22
2015/4/26	211406	3036	1.44	538	56	2442	1.16
2015/5/03	205698	554	0.27	347	51	156	0.08
2015/5/10	208653	519	0.25	384	49	86	0.04
2015/5/17	206371	1092	0.53	677	61	354	0.17
2015/5/24	162951	967	0.59	789	75	103	0.06
2015/5/31	72790	592	0.81	513	53	26	0.04
2015/6/07	78298	1317	1.68	1126	113	78	0.10

2015/6/14	85198	1047	1.23	681	85	281	0.33
2015/6/21	68848	676	0.98	422	76	178	0.26
2015/6/28	80548	535	0.66	368	39	128	0.16
2015/7/05	49218	146	0.30	125	3	18	0.04
2015/7/12	59131	70	0.12	45	2	23	0.04
2015/7/19	44165	52	0.12	39	2	11	0.02
2015/7/26	21596	30	0.14	25	1	4	0.04
2015/8/02	4476	4	0.09	2	0	2	0.04
2015/8/09	6377	2	0.03	1	0	1	0.02
2015/8/16	4574	3	0.07	3	0	0	0.00
2015/8/23	3937	2	0.05	2	0	0	0.00
2015/8/30	2831	2	0.07	0	0	2	0.07
2015/9/06	4711	6	0.13	2	1	3	0.06
2015/9/13	3182	4	0.13	0	1	3	0.09
2015/9/20	31908	22	0.72	6	0	16	0.05
2015/9/27	14162	11	0.08	4	0	7	0.05
TOTAL	1857775	11392	0.61	6365	700	4327	0.23

From ЗНО (EIT) Schedule table we could conclude that the most important weeks for “ЗНО” participants are:

1) April 20 –April 26 (as the exam on Ukrainian language and literature is on April 24) in which the number of tweets including the word ЗНО was the highest = 2,442 tweets;

2) May 11 – May 17 (as the announcement of the results of the exam on Ukrainian language and literature is on May,13) = 354 tweets;

3) June 8 – June 14 (as the exam on Math is on June 12) = 281 tweets;

4) June 15 – June 21 (as there are two major exams on this week: Biology – June 17, and History of Ukraine – June 19) = 178 tweets;

5) June 22 – June 28 (as the announcement of the results of the Math, English, Spanish, French, German exams on is on June 25) = 128 tweets;

6) June 29 – July 5 (as the announcement of the results of the Biology, History of Ukraine, Russian, and Physics is on July 3). We found an unexpectedly low number of tweets here = 18 tweets.

We see the intensification of Twitter activities on the above weeks, but the numbers of tweets using the word ЗНО are lower than we expected, except the April 20 –April 26 week. The total amount of tweets with either word “exam” (7,065) or word “ЗНО” (4,332) reaches 11,397. However, we can see that this portion constitutes around 0.6% of all 1,857,775 geotagged tweets sent in April – September 2015 in Ukraine. As “ЗНО” is significantly important only for those high-school students, who aim to enter prestigious universities in Ukraine, we admit, that our data cannot either prove or contradict the results of the previous research on Twitter activity, implying that the typical representative of Twitter users in Ukraine is a user characterized by young age and relatively high social and economic status (Kostenko, 2011; Pentina, Basmanova and Zhang, 2014).

Results on the proportion of tweets with word “EXAM” in Russian to those in Ukrainian

In our previous research we have found that in the period of April-August 2015 in the country as a whole, more than six Russian tweets are sent for every Ukrainian tweet (see.Pavliy and Lewis, 2015). Now we can state that the same tendency is seen with the tweets related to exams, Russian tweets heavily outnumber Ukrainian tweets even in time of “ЗНО” which is conducted in Ukrainian.

Table 3.　Proportion of tweets with word “EXAM” in Russian to those in Ukrainian

WEEK	Exam RU	Exam UK	% of exam UK to exam RU
2015/4/12	36	1	2.78
2015/4/19	230	31	13.48
2015/4/26	538	56	10.41
2015/5/3	347	51	14.70
2015/5/10	384	49	12.76
2015/5/17	677	61	9.01
2015/5/24	789	75	9.51
2015/5/31	513	53	10.33
2015/6/7	1126	113	10.04
2015/6/14	681	85	12.48
2015/6/21	422	76	18.01
2015/6/28	368	39	10.60
2015/7/5	125	3	2.40
2015/7/12	45	2	4.44
2015/7/19	39	2	5.13

2015/7/26	25	1	4.00
2015/8/2	2	0	0.00
2015/8/9	1	0	0.00
2015/8/16	3	0	0.00
2015/8/23	2	0	0.00
2015/8/30	0	0	0.00
2015/9/6	2	1	50.00
2015/9/13	0	1	n/a
2015/9/20	6	0	0.00
2015/9/27	4	0	0.00
Total:	6365	700	11.00

The ratio of Ukrainian to Russian tweets concerning exams is around 1:10, which means that in 2015 the tendency among high-school students and graduates in Ukraine in tweets related to examinations is even less than in general tweets (1:6). We understand the weaknesses of our research, such as inability to check the real age of a tweet sender, an access only to those tweets, who have geolocation, but still we can conclude that there is no visible trend among high-school graduates to prefer Ukrainian language over Russian in their daily life.

We can also state that the peak of tweeting activity is related to the examination period in ЗНО (EIT). However the percentage of tweets including the words ЗНО is less than 1 percent in all periods except the first week before the obligatory ЗНО exam of Ukrainian and literature (April, 24). As it was mentioned above, “ЗНО” is important only for those high-school students, who aim to enter prestigious universities, so relatively seldom usage of it cannot provide any evidence on a lower percentage of high-school students among Ukrainian internet users.

Conclusions

In our research we found that age related portion of tweets constitutes around 0.6% of all 1,857,775 geotagged tweets sent in April – September 2015 in Ukraine. With such an insignificant percent of high-school age users in our data we are unable to prove or refute the statement that the typical representative of Twitter users in Ukraine is a user characterized by young age. Moreover, we found it extremely difficult to identify the age of tweet senders with high accuracy, solely depending on the contents of their tweets.

As for the language in examination related tweets, we found that there is no visible tendency among young Ukrainians to prioritize Ukrainian language in their tweets concerning school or university examinations. Russian tweets concerning examinations outnumber Ukrainian ones with the proportion of Uk/Ru =1:10 (which is even less than the proportion (Uk/Ru = 1:6) in all geotagged tweets in the same period in general.

References

Bachmann, K., Lyubashenko, I. (2014) “The role of digital communication tools in mass mobilisation, information and propaganda” in *The Maidan uprising, separatism and foreign intervention : Ukraine's complex transition* Bachmann,K. and Lyubashenko, I. (eds.) Peter Lang.

Goban-Klas, T. (2014). EuroMaidan–Symbiosis of Political Protest and Media. *Open Europe: Cultural Dialogue Across Borders*, pp. 169-178.

Kostenko, N. (2011) “Information-Culture Styles in Russia and Ukraine.” *Sociological Research* 50 (4), pp. 57-86.

Mislove, A., Lehmann, S., Ahn, Y., Onnela, J., Rosenquist, J. (2011) “Understanding the demographics of Twitter users.” *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.*

Onuch, O. (2014) “The Middle Class Median Protester: EuroMaidan and Democratization in Ukraine”. *Journal of Democracy.* Vol.25 (3) pp. 44-51.

Onuch, O. (2015). EuroMaidan Protests in Ukraine: Social Media Versus Social Networks. *Problems of Post -Communism*, Vol. 62, pp.1-19.

Onuch, O. (2015) “Facebook Helped Me Do It” Understanding The EuroMaidan Protester ‘Tool-Kit’. *Studies in Ethnicity and Nationalism* 15 (1), pp. 170-184.

Osvita.ua. Визначено дати проведення тестів ЗНО у 2015 році.
http://osvita.ua/test/44778/ [Accessed on September 8, 2016]

Osvita.ua. МОН розпочинає підготовку до вступної кампанії http://osvita.ua/vnz/consultations/43954/ [Accessed on September 8, 2016]

Osvita.ua. Результати ЗНО (2015)
http://osvita.ua/test/rez_zno/ [Accessed on September 8, 2016]

Pavliy, B., Lewis, J. (2015) "Borders of Identity and Actual Language Use in Ukraine: An Analysis of Geotagged Tweets" Japanese Slavic and East European Studies, Vol.36, pp.77-97.

Pavliy, B., Lewis, J (2016) "The Performance of Twitter's Language Detection Algorithm and Google's Compact Language Detector on Language Detection in Ukrainian and Russian Tweets" *Bulletin of Toyama University of International Studies Faculty of Contemporary Society,* Vol.8, pp.99-106.

Pentina, I., Basmanova, O., Zhang, L. (2014) "A cross-national study of Twitter users' motivations and continuance intentions" *Journal of Marketing Communications* (doi:10.1080/13527266.2013.841273)

Ronzhyn, A. (2014) "The use of Facebook and Twitter during the 2013–2014 protests in Ukraine." In Proceedings of the European Conference on Social Media: ECSM, 2014, Reading, UK: Academic Conferences Ltd.

Ronzhyn, A. (2016) "Social Media Activism in Post-Euromaidan Ukrainian Politics and Civil Society" International Conference for E-Democracy and Open Government 2016, Conference Paper.

Sikorska, O., (2014) "Yandex Report: Twitter Usage in Ukraine" *Digital Eastfactor*
http://www.digitaleastfactor.com/yandex-report-twitter-usage-ukraine/ [Accessed on September 9, 2016]

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), article number: 7. (doi: 10.5153/sro.3001)

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), article number: e0115545.

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. doi:10.1371/journal.pone.0142209

Valenzuela, S., Valdimarsson, V., Egbunike, N., Fraser, M., Sey, A., Pallaev, T., Chachavalpongpun, P., Saka, E., Lyubashenko, I. (2014). “The Big Question: Have social media and/or smartphones disrupted life in your part of the world?” *World Policy Journal* 31(3), pp. 3-8.