The Performance of Twitter's Language Detection Algorithm and Google's Compact Language Detector on Language Detection in Ukrainian and Russian Tweets

Bogdan Pavliy and Jonathan Lewis

Abstract

In this paper we analyze the accuracy of Twitter's language detection algorithm and Google's Compact Language Detector in detecting and tagging the language of the tweets written in Ukrainian or Russian languages. The language recognition of the content of 4000 tweets by the two language detection tools is compared with the language identification by bilingual native speakers of Ukrainian and Russian. We discuss some difficulties in identifying a given tweet's language; some difficulties are specific to Ukrainian and Russian while others are due to the Twitter medium. We show that the performance of the Google algorithm can be improved by cleaning the tweets before running the algorithm.

Keywords: language, social networks, Twitter, Ukrainian, Russian

Introduction

Since the early 2000s academic researchers have started to use social networks and online social media for their studies. Among most popular social media, scholars have focused on the microblogging service Twitter because of the ease of access to the data. Twitter allows users to publish their location at the time of posting. For privacy reasons, the user is required to opt in to location publishing; as a result, only about one percent of tweets are geotagged (Jurgens et al., 2015, Johnson et al., 2016). Twitter's Streaming API lets researchers request all tweets geotagged within a given area and maintain a continuous collection of tweets sent from a certain territory. The accessibility of Twitter data has stimulated research using the microblogging service as a social sensor for examining diverse aspects of human behavior such as political debate (Conover et al., 2011), rumors following natural disasters such as hurricanes (Kogan et al.2015) or earthquakes (Takayasu et al., 2015), and reactions during sporting events (Takeichi et al., 2014).

Investigation of the linguistic aspects of communication on Twitter is facilitated by the existence of language detection algorithms that permit the automatic identification of the language used in the text of tweets. Twitter runs its own language detection algorithm on each tweet and provides the result—a single language tag for each tweet. However, the algorithm is not publicly available and hence cannot be used to recognize the language of other content including other popular social

networking services such as Facebook, Google+, LinkedIn, etc. In cases where Twitter's own algorithm proves insufficiently accurate or where recognition of multiple languages is required, other algorithms, such as Google's Compact Language Detector, are available.

As part of an ongoing study on language preferences of Twitter users in Ukraine, it was necessary to assess the accuracy of the Twitter's Ukrainian or Russian language detection. We therefore considered the possibility to use Google's Compact Language Detector instead of Twitter's language detection algorithm if it proves to be more accurate. To check the level of accuracy of both systems, we decided to ask native Ukrainian-Russian bilingual speakers to identify the language of the tweets and then compare their results with the language identification by Twitter's and Google's language detection algorithms.

Data Collection and Initial Cleaning

We collected geotagged tweets sent from the territory of Ukraine (including Crimea) from the Twitter Streaming API between April 11 and September 15, 2015. This was achieved by writing a Python script using the tweepy library, which established an open connection to the Twitter Streaming API and specified the geo-coordinates of a bounding box that contained the territory of Ukraine (including Crimea). Whenever a geotagged tweet was sent from within the bounding box, our program received it and stored it in a PostGreSQL database. We then excluded tweets sent from areas in our bounding box that were outside the territory of Ukraine. We also excluded tweets generated by the location service Foursquare that merely included text information about the user's location.

Analysis of the Accuracy of Twitter and Google Language Detection

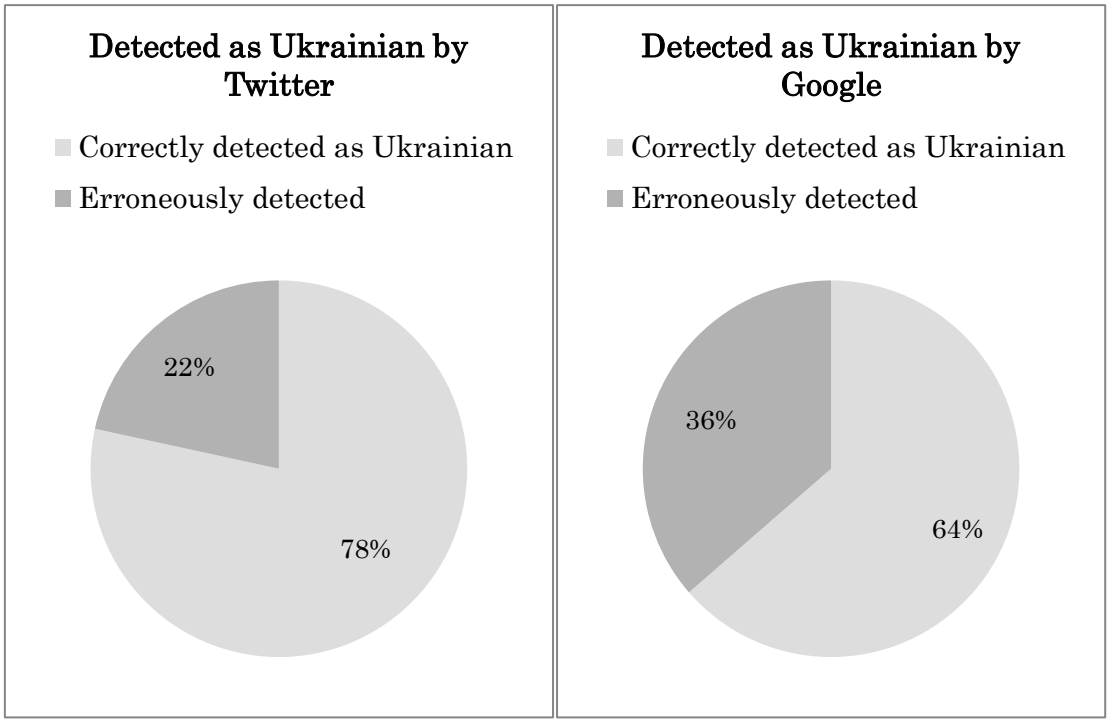
To provide a sufficient analysis of the accuracy of Twitter's and Googles language identification, we selected at random 2000 tweets recognized by Twitter language detection algorithm as written in Ukrainian and 2000 tweets recognized as written in Russian. Then we had them checked by bilingual native speakers of Ukrainian and Russian languages and compared the results, which are summarized in Table 1.

Table 1.

Tweets language detected as UK by Twitter	Tweets language detected as RU by Twitter	Both Twitter and native speaker detected as UK	Both Twitter and native speaker detected as RU	Both Google and native speaker detected as UK	Both Google and native speaker detected as RU
2000	0	1568	0	1272	0
0	2000	0	1846	0	1337

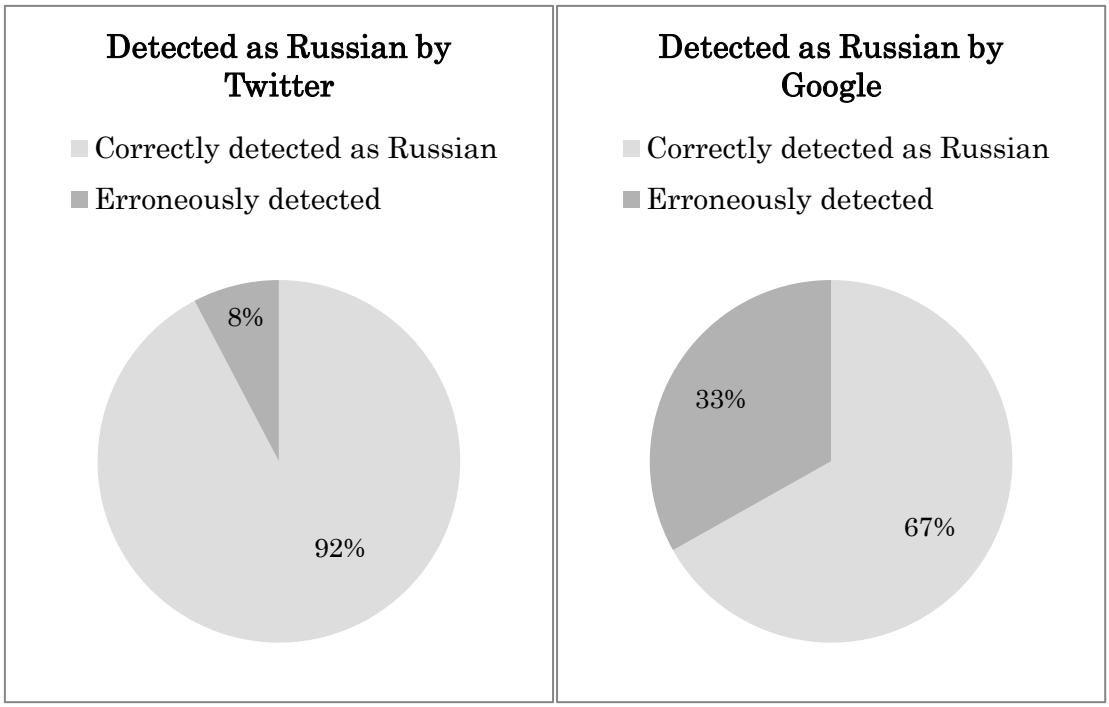
As we can see, the performance of Twitter's language detection algorithm is considerably better than Google's. In case of Ukrainian language detection, the correctness of Twitter is 78%, while initial level of Google's correctness is 64% (Fig.1)

Fig.1



In case of Russian language detection, the performance of Twitter is even better: 92%, while initial level of Google's correctness is 67% (Fig.2)

Fig.2



Clearly, Twitter's algorithm achieves a higher level of accuracy than Google's, especially in identifying tweets written in Russian. However, we investigated the variations in language tags and found that Google's Compact Language Detector often identifies the language of the tweets as NONE. Hence it would not be correct to assert that Twitter's performance on language recognition always surpasses Google's. In our batch of 4000 tweets Google detected as NONE 1510 tweets, where 807 tweets were detected by Twitter algorithm as UK Ukrainian and 703 tweets as Russian. As the number of such tweets exceeds one third of all tweets, there is a need to analyze their content and find ways to improve the language detection in case of using Google's algorithm.

Tweets Tagged by Google as NONE

Having analyzed the content of tweets tagged by Google as NONE, we came to the following conclusions:

1. In general, most of the tweets in the NONE category contained very short messages, where even native speakers sometimes had difficulty understanding the meaning of the tweets.

2. Among the tweets in NONE category, there were tweets written in surzhyk (a mixture of both Ukrainian and Russian), and in some cases language identification was problematic for native speakers of both languages.

3. The similarity of Ukrainian and Russian expressions is a major problem in detecting the language for native speakers. As some expressions are identical in both languages the expressions alone cannot be detected as either Ukrainian or Russian, so even the native speakers decided to mark them as NONE e.g. “Христос Воскрес!” = Jesus Has Risen! (identified by native speakers as NONE))

4. Use of emoticons, hashtags, abbreviations (e.g. Шалено 🙈🙈🙈🤾🔥🔥□🚀🚀🚀😈😈😈😈 #бумбокс = Crazy 🙈🙈🙈🤾🔥🔥□🚀🚀🚀😈😈😈😈 #boombox (identified by native speakers as Ukrainian))

5. Expressions with no meaning, mimicking sounds (e.g.Бам Бам Бам Бам= Bum Bum Bum Bum (identified by native speakers as NONE))

6. Mixing languages by using English words inside of Ukrainian or Russian phrase (e.g. “Де твій Online коли ти так потрібна” = Where is your Online when I need you so (identified by native speakers as Ukrainian))

7. Writing English expressions in Cyrillic alphabet (e.g. “май фейфоріт піца” = My favorite pizza (identified by native speakers as NONE)).

8. Use of slang, spelling mistakes, ungrammatical writing or compressed writing (“мала,з др” = Happy Birthday, baby (identified by native speakers as Ukrainian))

9. Mixture of Latin and Cyrillic letters in one word (e.g. “Sportик” (identified by native speaker as NONE).

10. Repetition of some letter(s) in emotional expressions (e.g. “ТИ СЕРЙООЗНОО” = ARE YOU SEERIOOOUS (identified by native speakers as Ukrainian)).

It is highly probable that the above problems caused more than one third of the tweets from our batch to be recognized as NONE by Google’s algorithm. Consequently, we decided to perform cleaning of the tweets content and discuss how we could further improve the accuracy, based on the results of language detection by native speakers.

Process of Cleaning and the Results of Cleaning

To perform the cleaning of the 4000 tweets, we wrote and ran a Python script which removed URLs, @usernames and #hashtags from the text of tweets before running the Google language detection algorithm.

After careful selection and analyses of the results (both positive and negative) of cleaning, we found that running the cleaning script had five effects for tweets identified by Twitter as either

Ukrainian or Russian and by Google as NONE (see Table 2 for details):

1. *Changed incorrectly*

The new language tag does not match native speakers' detection.

2. *Changed to Ukrainian correctly*

The new language tag matches native speakers' language detection as Ukrainian.

3. *Changed to Russian correctly*

The new language tag matches native speakers' language detection as Russian.

4. *Erroneous change from NONE to some language*

Google's initial identification of the tweet's language was correct and matched native speakers' detection as NONE, but after cleaning Google erroneously identified the language as UK, RU or some other language.

5. An improvement of the recognition of tweets written in the Belarusian language. Three tweets initially tagged as NONE after cleaning were correctly recognized as Belarusian. However, as we target only on Russian and Ukrainian languages in this study, we will not discuss this further here.

Table 2.

Type of change / Detected by Twitter	**1) Changed incorrectly** (do not match native speakers' detection)	**2) Changed to Ukrainian correctly** (match native speakers' language detection as Ukrainian)	**3) Changed to Russian correctly** (match native speakers' language detection as Russian)	**4) Erroneous change from NONE** (matched native speakers' NONE detection but was erroneously given some language tag after cleaning)	**TOTAL**
as Ukrainian	**59**	**204**	**42**	**14**	**319**
as Russian	**9**	**1**	**258**	**8**	**277**
Total	**68**	**205**	**300**	**22**	**596**

From this data we can conclude that, based on native speakers' recognition of the language of each tweet, we got better results in language identification by Google after cleaning. From our batch of 4000 tweets in 1510 tweets that were tagged as NONE after cleaning recognition improved for 508 tweets (205 Ukrainian, 300 Russian, 3 Belarusian), while negative changes caused by recognition errors due to cleaning happened only for 22 tweets.

Conclusion

The results at the first stage of language identification by Twitter and Google showed that Twitter's performance, especially in Russian language recognition is better that Google's and without cleaning the difference is immense. This is probably to be expected: Twitter will have optimized its algorithm for shorter texts and probably also ignores URLs, hashtags and usernames when identifying the language of each tweet. However after proper cleaning of the content of the tweets, we can improve the Google's recognition and conclude (as it was suggested before) that both Twitter's and Google's language detection systems can be fairly accurate in Ukrainian and Russian language recognition. While working on that paper, we found some other distractors in forms of set phrases or expressions (e.g. Найден новый адрес = New address was found) which can be the subjects for further cleaning steps. At this stage Twitter still seems to be more accurate than Google in recognizing Ukrainian and Russian languages in tweets, even after cleaning data before running the Google algorithm. However there may be further scope for cleaning data to improve the performance of the Google's algorithm.

Acknowledgments

We thank NIFTY Corporation for the use of the C4SA cloud hosting service for data collection. We also thank the developers of the following open source software packages used in this research: PostGreSQL and Python.

References

Conover, M. D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Flammini, A., and Menczer, F. (2011) "Political Polarization on Twitter." Fifth International AAAI Conference on Weblogs and Social Media. Available at:
http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/index.

Johnson, I., Sengupta, S., Schöning, J., Hecht, B. (2016) The Geography and Importance of Localness in Geotagged Social Media. Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2016), New York: ACM Press

Jurgens, D., Finethy, T., McCorriston, J., Xu, Y., Ruths. D. (2015) Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice" ICSCW

Kogan,M., Palen,L., Anderson,K., (2015) Think Local, Retweet Global: Retweeting by the Geographically-Vulnerable during Hurricane Sandy. CSCW 15, Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, ACM Press,pp.981-993 doi:10.1145/2675133.2675218

Takayasu, M., Sato, K., Sano, Y., Yamada, K., Miura, W., and Takayasu, H. (2015) "Rumor Diffusion and Convergence during the 3.11 Earthquake: A Twitter Case Study," PLoS ONE 10 (4): e0121443. doi:10.1371/journal.pone.0121443.

Takeichi, Y., Sasahara, K., Suzuki, R., and Arita, T. (2014) "Twitter as Social Sensor: Dynamics and Structure in Major Sporting Events," ALIFE 14: Proceedings of the Fourteenth International Conference on the Synthesis and Simulation of Living Systems. doi:10.7551/978-0-262-32621-6-ch126.