Who are Birds of the Same Feather? Epistemic Communities in the EU Twittersphere

Oul Klara Han
Graduate School of
East Asian Studies
Free University Berlin
hanoul@gmail.com

Suin Kim
School of Computing
KAIST
suin.kim@kaist.ac.kr

Camille Roth
Centre Marc Bloch
CNRS/MAEE/HU
roth@cmb.hu-berlin.de

Abstract

We are increasingly faced with pan-EU issues that polarize and mobilize, such as the Grexit or the refugee crisis. The epistemic connections regarding the perception of pan-EU issues among EU member countries are still under-explored: are discussions on EU issues determined by national barriers (e.g. language), or do they instead divide into epistemic communities that span across multilingual barriers? Our paper applies this question on an online medium by conducting multilingual twitter analysis on cross-national links between epistemic communities, thereby testing the notion of an EU Twittersphere. Our aim is to describe a typology of various topic-centered user networks: in other words, do some issues correspond to different (pan-EU) network structures?

Keywords: epistemic communities, multilingual LDA, topic modeling, Twitter, EU

The notion of epistemic community (Roth and Bourgine, 2006) is relevant to study the conceptual possibility of a European public sphere that forms around shared issues within and across member states (Lietz et al., 2014; Verdun, 1999). For this purpose, it is essential to look beyond natural national barriers such as language. It is already known that online public space actors assume structural roles within monolingual/mononational epistemic communities (for instance in blogspace, see Cardon et al., 2011; see also Poell and Darmoni, 2012). Also, clustering patterns of online communities affect the spread of information flows (Weng et al., 2013). Thus, we contend that the understanding of clustering patterns in the EU Twittersphere would shed light on the causal mechanism of opinion-forming and disseminating by the EUs political actors who are active on Twitter. By applying the concept of epistemic communities to the EU Twittersphere, we aim at observing the configuration of multi-lingual arenas in social media. This would additionally be a methodological advancement over more traditional content analysis methods that were used in political science for similar questions (Machill et al., 2006).

In the big picture, the EU Twittersphere exists alongside EU politics, but possesses its own distinctive dynamics of information dissemination and clustering. While the EU Twittersphere communicates and disseminates, EU politics is the arena where higher-level policies and decisions materialize. In this sense, we can say that EU politics is a superstructure for the EU Twittersphere. At the same time, the EU Twittersphere is an openly accessible, individually maneuverable, and quick-paced platform for spreading knowledge. Thus, the naturally constraining effects of language and national barriers may prove less influential in forming epistemic communities in the EU Twittersphere. In order to find out and differentiate how the realm of the EU Twittersphere actually behaves in relation to EU issues, we require methods that deal with multilingual and massive corpora of tweets. Our approach has the potential to enrich many of the existing studies. Firstly, it enriches the perspective of structural studies, or structural studies combined with a binary indicator (e.g. pro-/anti-European (Hoffmann, 2014) or the now classical "Divided they Blog" (Adamic and Glance, 2005) and subse-

quent similar studies). Secondly, this approach enriches socio-semantic studies on several countries featuring a manual content analysis of a digital public space defined around a single language (such asEtling et al., 2009).

Data

The dataset consists of multilingual tweets, crawled from a seed list of 4,500 Twitter users who were manually identified as politically relevant users [from the study of Maireder, Schlogl, Schutz, Karwautz, Waldheim, 2014, and kindly provided by Axel Maireder, whom we gratefully thank]. A large proportion (95%) of these accounts are still active as of today, and we assume that they are representative of the current public space for the respective countries (mainly from France, Germany, Great Britain and Ireland). For each user in the list we crawled the 3,200 most recent tweets. We then detected the language of each tweet using *chromium language detector 2*. We also filtered the dataset to only include the tweets published from September 2014 to further reduce the potential bias stemming from the heterogeneous tweet publication activity across users. About 52% of the users we crawled have their first publicly available tweet published later than September 2014. After the filtering, the number of tweets for each language yields the highest volumes for English, French, and German comprises 94.3% of all dataset (3.73M 40.3%, 3.52M 38.0%, and 1.48M 16.0% each), out of the 9.26M tweets from 81 languages. Hence, we focus our analysis on detecting links between the corpora of above three languages. We finally extracted the mention network restrained to the original seed list, which yields 651K edges (average 148 edges per user).

Method

We carry out a socio-semantic analysis on the content of prominent issues across EU Twitter users. In a nutshell, we need to define and then study user networks corresponding to distinct issues. To this end, we first use a topic modeling framework to (1) automatically discover the set of topics from the un-annotated dataset and (2) (soft-) cluster each tweet or each Twitter user into the discovered topics. Latent Dirichlet allocation (Blei et al., 2003) is the widely used topic modeling framework. Based on the topic modeling results, we then derive topic-based networks. We observe sub-networks that strongly hint at epistemic communities that transcend language barriers across EU member states.

References

[1] Adamic, L.A., Glance, N., 2005. The political blogosphere and the 2004 US election: divided they blog, in: Proceedings of the 3rd International Workshop on Link Discovery. ACM, pp. 3643.

[2] Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 9931022.

[3] Cardon, D., Fouetillou, G., Roth, C., 2011. Two Paths of Glory-Structural Positions and Trajectories of Websites within Their Topical Territory., in: ICWSM.

[4] Etling, B., Kelly, J., Faris, R., Palfrey, J., 2009. Mapping the Arabic blogosphere: politics, culture, and dissent. Internet & Democracy Project, Berkman Center for Internet & Society.

[5] Hoffmann, I., 2014. spotlight europe 2014/02, Mai 2014: Im Netz der Populisten. Bertelsmann Stiftung.

[6] Machill, M., Beiler, M., Fischer, C., 2006. Europe-Topics in Europes Media The Debate about the European Public Sphere: A Meta-Analysis of Media Content Analyses. Eur. J. Commun. 21, 5788. doi:10.1177/0267323106060989

[7] Maireder, A., Schlögl, S., Schütz, F., Karwautz, M., Waldheim, C., 2014. The European Political Twittersphere [WWW Document]. Univ. Wien GfK. URL http://www.gfk.com/twitter

[8] Poell, T., Darmoni, K., 2012. Twitter as a multilingual space: The articulation of the Tunisian revolution through #sidibouzid. NECSUS Eur. J. Media Stud. 1, 1434.

[9] Roth, C., Bourgine, P., 2006. Lattice-based dynamic and overlapping taxonomies: The case of epistemic communities. Scientometrics 69, 429447.

[10] Weng, L., Menczer, F., Ahn, Y.-Y., 2013. Virality prediction and community structure in social networks. Sci. Rep. 3.