25

Analysing the social fingerprints of pro-independence movements

Arkaitz Zubiaga, University of Warwick

Researchers increasingly turn to social media to analyse the online fingerprints produced by offline political movements. Social media such as Twitter have been exploited to learn about society in events such as elections, major protests and other crises and emergencies. However, the analysis of smaller, more obscure communities (with respect to Twitter, but which have an otherwise salient presence), has been studied to a lesser extent. This is the case of pro-independence movements where a smaller part of an existing country makes claims for its independence from the rest of the country. There might be many pro-independence supporters within that region, but that does not always suffice to make it to Twitter's list of trending topics and/or to reach a broad share of Twitter users, which is however crucial to ensure that the community is visible to others.

Credit: Andreas Eldh/Flickr/CC BY 2.0

My initial attempts at studying pro-independence movements on social media have focused on the cases of the Basque Country and Catalonia, two regions which, as of today, are mostly part of Spain and to a lesser extent of France, but have historically had different cultural and linguistic backgrounds compared to the rest of Spain. The case of these two lands is somewhat different to that of Scotland. While the latter did have an opportunity to vote for leaving or staying in the United Kingdom, the former are not entitled to run a referendum, as Spain deems it unconstitutional. Due to this, Basques and Catalans have organised demonstrations protesting for their 'right to decide'. Here I look at two such demonstrations, both in the form of human chains, one 300 miles long in Catalonia in 2013 (Via Catalana, or 'the Catalan Way'), and the other a 76 miles long in the Basque Country in 2014 (Gure Esku Dago, or 'It's in Our Hands'). In both cases, I collected tweets using Twitter's streaming API both on the day of the demonstration as well as during the week preceding it. I used the main keyword in each case to retrieve tweets relevant to the event in question: #gureeskudago for the Basque demonstration and #viacatalana for the Catalan demonstration.

Table 1 shows the top hashtags that people used in each case. We can observe that other related hashtags emerge, such as #GureEskuBidea (from Basque 'the way is in our hands') and #FemVia (from Catalan 'we make the way'), both in their respective languages. Besides those, a number of hashtags on top of the list are in English, such as #BasquesDecide, #CatalansWantToVote, #Up4Freedom and

#CatalanWay, while there are no hashtags in Spanish. This shows how both communities are using the English language to try and reach out to other communities and users from other countries so as to increase the visibility of the demonstrations and of their demands. The high popularity of the #CatalanWantToVote hashtag in the Basque demonstration shows in turn the solidarity of these two communities, both of which are asking Spain to let them determine by vote their future status.

Table 1: Top hashtags in the Basque and Catalan demonstrations

#gureeskudago (Basque)		#viacatalana (Catalan)	
#GureEskuDago	34,542	#ViaCatalana	238,343
#BasquesDecide	17,463	#Croquetes	53,276
#CatalansWantToVote	1,077	#CatalanWay	29,353
#GureEskuBideaE8	1,028	#11s2013	26,777
#GureEskuBidea	826	#Diada	11,270
#Up4Freedom	686	#Independència	7,982
#Gizakatea	429	#FemVia	7,417
#BasqueDecide	409	#ViaCatalanaTV3	7,260
#EuskalHerria	332	#Independencia	5,810
#NiBanoa	326	#Catalunya	4,991

While looking at hashtags provides valuable insight into these events, one might want to dig deeper into this data by, for instance, looking at the content of the tweets. The largeness of the data sets, which would make the work of a researcher sifting through the data cumbersome, requires the use of automated methods for textual analysis and natural language processing (NLP). However, having to perform the analyses for these communities, each of which has its own language, leads to the following two additional issues to deal with:

1. While Twitter does tell us what language each tweet is written in through its API, this feature is limited to some major languages. Twitter does not identify tweets written in Basque or Catalan at the time of this writing. For a tweet written in Basque or Catalan, Twitter will instead tell us that it is written in another, often unrelated language. Table 2 shows the statistics of languages reported by Twitter for the tweets in the two collections. Spanish is listed

in both cases as the top language, which is unlikely to be true. Moreover, the fact that Bosnian, Indonesian and Tagalog are so popular in the case of the Basque demonstration is implausible pointing to the limitations of Twitter's language identification algorithm. While the majority of tweets will most likely be in Basque and Catalan, Twitter will never know that, and we get instead other sets of languages from its API. The accurate identification of the language of a tweet, especially in the case of languages spoken by smaller groups of people, is still an open research task which we have been recently working on through the organisation of a workshop and shared task called TweetLID.

2. The automated linguistic analysis of tweets is often performed with well-known tools for NLP such as GATE, psychologically informed word counts analysis tools such as LIWC, or bespoke solutions for Twitter such as Tweet NLP. However, these tools tend to be crafted to perform well for a few major languages and hence, they leave much to be desired when they are applied to new languages. The development of both linguistic resources and NLP tools suitable for these languages will be crucial for a successful analysis of this kind of data.

Table 2: Twitter's automated language identification results

#gureeskudago (Basque)		#viacatalana (Catalan)	
Spanish	10,518	Spanish	160,771
Bosnia	3,973	Portuguese	24,314
Indonesian	3,936	French	20,556
Tagalog	3,893	Italian	19,266
Croatian	3,853	English	17,225
Estonian	3,552	Romanian	15,233
English	2,786	Indonesian	4,419
Romanian	1,626	Slovak	3.622
Finnish	799	Welsh	2,859
Portuguese	578	Tagalog	2,763

I am confident that within the next few years we will see an increasing body of research studying political issues around other such smaller

communities from around the world and we will be developing more sophisticated techniques for textual analysis of tweets for lesser-resourced, minority languages.

Arkaitz Zubiaga is a post-doctoral research fellow at the University of Warwick, currently involved in the PHEME FP7 project on the study of social media rumours. His research revolves around the study of the spread of news and events through social media and especially the role of citizen journalists in news reporting. Some of his recent research has dealt with the curation, verification and classification of newsworthy information shared on social media involving computational research in fields like text mining, natural language processing and social computing, but also including an interdisciplinary perspective from the social sciences such as journalism and psychology. He can be found tweeting as @arkaitz.